Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

Supplementary Material

1. Performance comparison to reference models

a. Models

We compared the performance of the APANN model with three commonly used models 1) linear regression with Lasso regularizer (LR_L1) (Casanova et al., 2013; Westman, Muehlboeck, & Simmons, 2012), 2) support vector regression (SVR) (Davatzikos, Bhatt, Shaw, Batmanghelich, & Trojanowski, 2011; Hinrichs, Singh, Xu, Johnson, & Alzheimers Disease Neuroimaging Initiative, 2011; Vemuri et al., 2008; Zhang, Shen, & Alzheimer's Disease Neuroimaging Initiative, 2012), and 3) random forest regression (RFR) (Gray et al., 2013). Separate instances of these baseline models were trained for MMSE and ADAS-13 prediction tasks. Separate instances of these models were also trained to compare performance of each each input, namely: 1) HC, 2) CT, and 3) HC+CT. The input features from each individual modality for the three baseline models were as follows:

1. HC: 2 continuous variables representing left and right hippocampal volumes
2. CT: 78 continuous variables representing thickness values based on AAL atlas ROIs (Tzourio-Mazoyer 2002)

The difference in input feature sets for the baseline models was prompted by the use of anatomically driven, low-dimensional features in many instances in the relevant literature (Suk, Lee, Shen, & Alzheimer's Disease Neuroimaging Initiative, 2015; Zhang et al., 2012). Moreover, we also investigated high-dimensional HC and CT input choices (identical as of APANN model) for the baseline models. However, the baseline models considerably underperformed with high-dimensional input compared to the input choice of low-dimensional features. The input values for LR_L1 and SVR models were preprocessed with an additional step in which data were mean centered and feature-wise scaled to unit variance. All the baseline models were implemented using scikit-learn toolbox (http://scikit-learn.org/stable/index.html).

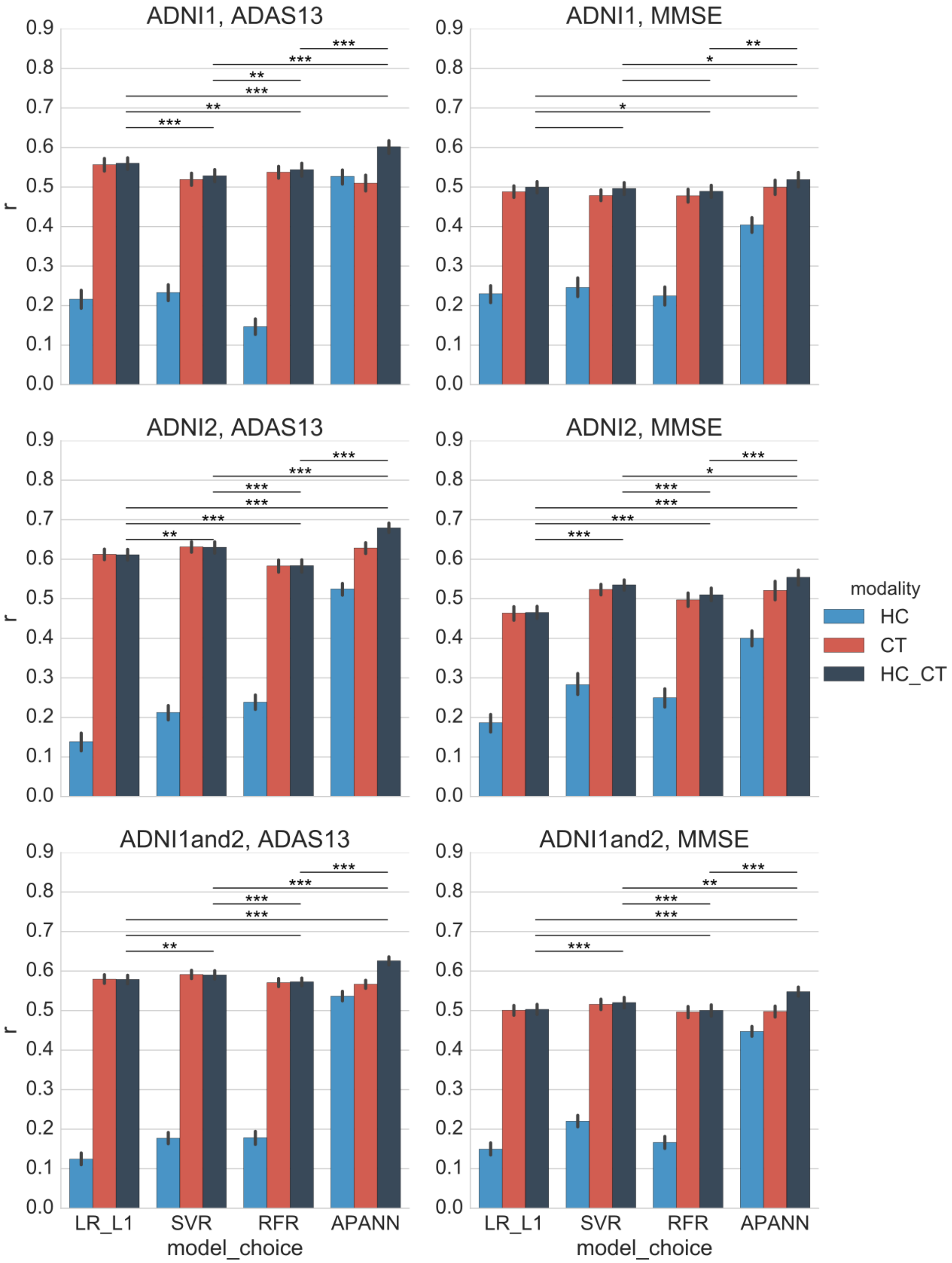
b. Results

The correlation performance comparison between APANN and reference models for both tasks and three experiments is shown below in Fig. S1. All correlation and rmse values are also tabulated in Table 1 in the manuscript. Correlation performance of baseline models with high-dimension input (identical as of APANN model) is shown in Fig. S2.

Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

Online appendices are unedited and posted as supplied by the authors.



Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

Figure S1: Performance of four models subject to individual and combined input modalities. The correlation values are averaged over 10 rounds of 10-folds. All models were trained with a nested-inner loop that searched for optimal hyperparameters. The statistical significance[1] annotations for pairwise method comparisons are as follows: *: p<0.05, **: p<0.01, ***: p<0.001. Plots only show statistical comparisons for HC+CT modality.

[1]https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ttest_rel.html

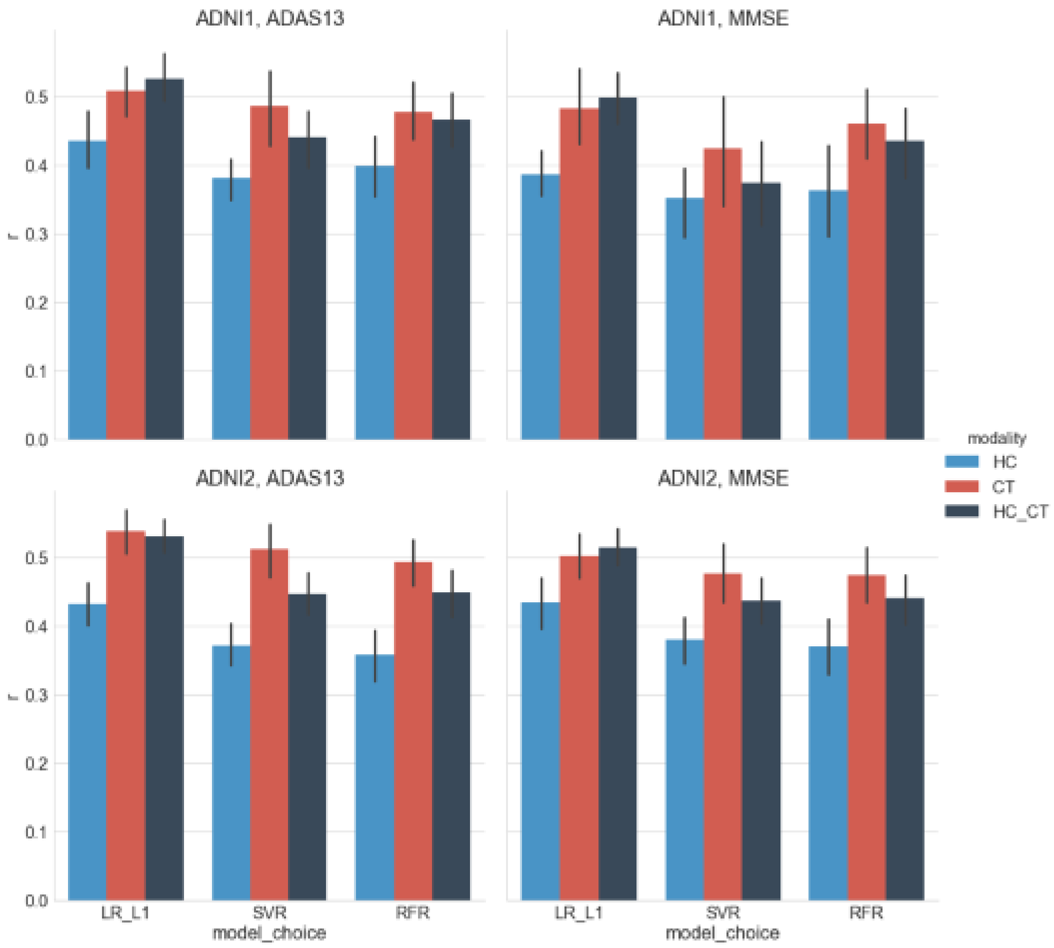


Figure S2: Correlation performance of LR_L1, SVR, and RFR models with high-dimensional input comprising HC: 32557 voxel-wise hippocampal features and CT: 686 thickness values from ROIs.

c. Discussion

Results from all three experiments indicate that APANN model offers better predictive performance with HC and HC+CT inputs. Specifically for the HC input modality, we see substantial improvement with

APANN model utilizing voxel-wise information from segmented hippocampal masks. We note that this performance gain is attributed to 1) added information from the voxel-wise input compared to two volumetric measures, and 2) the computational capacity of APANN to extract useful features from this voxel-wise input compared to the baseline models. As described earlier, the baseline models do show improved performance with the voxel-wise input, however, the gains are smaller compared to the APANN model. In comparison, CT input modality, when used independently, does not offer improvement with APANN model. For ADAS-13 prediction task, the baseline models outperform the APANN model in Experiment 1 and 3, and offer similar performance in Experiment 2 when comparing performance with CT input alone. However, the HC+CT input to APANN model offers significantly higher performance improvement over baseline models across all three experiments.

Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

2. Empirical sampling: standardization across modalities

Since we use independent procedures for augmenting number of samples for the two input modalities (HC and CT), each subject may end up with different number of empirical samples from each input. This raises an issue during training models with combined input from both modalities (HC+CT). Additionally, within each modality, two subjects may have drastically different number of empirical samples, which can result in biased training. In order to avoid these issues, we standardize the empirical sample sizes by enforcing following constraints: 1) the total number of samples per subject need to be equal across modalities (# of HC samples = # of CT samples for i^{th} subject); and 2) the number of samples per subject needs to be similar in number across subjects (# of samples for i^{th} subject ~ # of samples for j^{th} subject). Based on these constraints, the samples are randomly chosen from the available pool of empirical samples for a given subject. At the end of this process, the resultant augmented training set sample size was approximately 35 times the number of subjects in the cohort.

Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

3. Computational resource requirements

The ANN models were trained using NVIDIA GPU GeForce GTX TITAN X and Caffe deep learning framework (http://caffe.berkeleyvision.org/). Training durations for a single model for each input modality averaged over the different hyperparameter combinations (number of hidden nodes, learning rate etc.) are as follows: 1) CT (input dimensionality: 686): 4 minutes (Experiment 1 and 2), 7 minutes (Experiment 3), 2) HC (input dimensionality: 32557): 8 minutes (Experiment 1 and 2), 15 minutes (Experiment 3), 3) HC+CT (input dimensionality: 33243): 11 minutes (Experiment 1 and 2) 20 minutes (Experiment 3).

Appendix 1 to Bhagwat N, Pipitone J, Voineskos AN, et al. An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *J Psychiatry Neurosci* 2019.

DOI: 10.1503/jpn.180016

4. Performance bias in combined ADNI1and2 cohort

In Experiment 3, we used the pooled ADNI1 and ADNI2 datasets during cross-validation. Therefore models were trained and tested on mixtured of ADNI1 and ADNI2 subjects. In order to evaluate if these trained models show any dataset bias in their predictive performance, we stratified the test performance based on subject-dataset membership (ADNI1 vs. ADNI2). Then the performance bias was computed as follows:

$$\square\square\square\square\,(\square\square\square\square\square\square\square\square\square\square\square) = \frac{\square\square\square\square(|\square\_\square\square\square\square 2 - \square\_\square\square\square\square 1|)\ )}{\square\square\square\square(\square\_\square\square\square\square 1\square\square\square 2)}$$

Based on this metric, a model exhibits a high bias if there is a large difference between performances of ADNI1 and ADNI2 splits (i.e. models performing well on only single dataset). The results show that APANN has the smallest performance bias towards any particular dataset compared to other models for both ADAS13 and MMSE tasks.

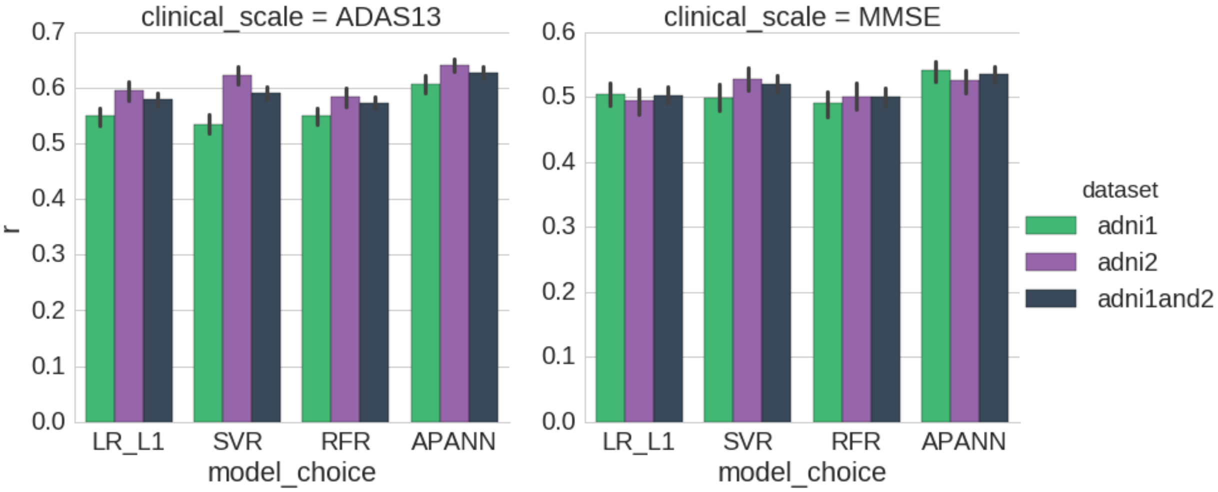


Figure S4: Correlation (r) performance of four models for the HC+CT input in Experiment 3 split by subject-dataset membership. In contrast with Experiment 1 and 2, all models were cross-validated on pooled ADNI1+ADNI2 subjects. The stratification of results offers insight into performance biases that may have caused by the dataset membership itself (eg.: high prediction performance for only ADNI1 subjects). APANN model outperforms other models in dataset-wise comparisons for both clinical scales and exhibits low bias towards single dataset.

Table S4: Bias measures for Experiment 3.

	LR_L1	SVR	RFR	APANN

ADAS13	0.1735	0.1958	0.1773	0.1284
MMSE	0.2129	0.2150	0.2320	0.1991

References

Casanova, R., Hsu, F.-C., Sink, K. M., Rapp, S. R., Williamson, J. D., Resnick, S. M., … for the Alzheimer’s Disease Neuroimaging Initiative. (2013). Alzheimer’s Disease Risk Assessment Using Large-Scale Machine Learning Methods. *PloS One*, *8*(11), e77949.

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N., & Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, *32*(12), 2322.e19–e27.

Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., & Alzheimer’s Disease Neuroimaging Initiative. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer’s disease. *NeuroImage*, *65*, 167–175.

Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., & Alzheimers Disease Neuroimaging Initiative. (2011). Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. *NeuroImage*, *55*(2), 574–589.

Suk, H.-I., Lee, S.-W., Shen, D., & Alzheimer’s Disease Neuroimaging Initiative. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Structure & Function*, *220*(2), 841–859.

Vemuri, P., Gunter, J. L., Senjem, M. L., Whitwell, J. L., Kantarci, K., Knopman, D. S., … Jack, C. R., Jr. (2008). Alzheimer’s disease diagnosis in individual subjects using structural MR images: validation studies. *NeuroImage*, *39*(3), 1186–1197.

Westman, E., Muehlboeck, J.-S., & Simmons, A. (2012). Combining MRI and CSF measures for classification of Alzheimer’s disease and prediction of mild cognitive impairment conversion. *NeuroImage*, *62*(1), 229–238.

Zhang, D., Shen, D., & Alzheimer’s Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer’s

DOI: 10.1503/jpn.180016

disease. *NeuroImage*, *59*(2), 895–907.