Predicting Alzheimer's disease progression using deep recurrent neural networks

Minh Nguyen[1,2], Tong He[1,2], Lijun An[1,2], Daniel C. Alexander[3], Jiashi Feng[1], B.T. Thomas Yeo[1,2,4,5,6] for the Alzheimer's Disease Neuroimaging Initiative*

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, [2]Clinical Imaging Research Centre, N.1 Institute for Health and Memory Networks Program, National University of Singapore, Singapore [3]Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK [4]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA [5]Centre for Cognitive Neuroscience, Duke-NUS Medical School, Singapore [6]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore

Address correspondence to:

B.T. Thomas Yeo
ECE, CIRC, N.1 & MNP
National University of Singapore
Email: thomas.yeo@nus.edu.sg

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

Early identification of individuals at risk of developing Alzheimer's disease (AD) dementia is important for developing disease-modifying therapies. In this study, given multimodal AD markers and clinical diagnosis of an individual from one or more timepoints, we seek to predict the clinical diagnosis, cognition and ventricular volume of the individual for every month (indefinitely) into the future. We proposed a recurrent neural network (RNN) model and applied it to data from The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge, comprising longitudinal data of 1677 participants (Marinescu et al. 2018) from the Alzheimer's Disease Neuroimaging Initiative (ADNI). We compared the performance of the RNN model and three baseline algorithms up to 6 years into the future. Most previous work on predicting AD progression ignore the issue of missing data, which is a prevalent issue in longitudinal data. Here, we explored three different strategies to handle missing data. Two of the strategies treated the missing data as a "preprocessing" issue, by imputing the missing data using the previous timepoint ("forward filling") or linear interpolation ("linear filling). The third strategy utilized the RNN model itself to fill in the missing data both during training and testing ("model filling"). Our analyses suggest that the RNN with "model filling" was better than baseline algorithms, including support vector machine/regression and linear state space (LSS) models. However, there was no statistical difference between the RNN and LSS for predicting cognition and ventricular volume. Importantly, although the training procedure utilized longitudinal data, we found that the trained RNN model exhibited similar performance, when using only 1 input timepoint or 4 input timepoints, suggesting that our approach might work well with just cross-sectional data. An earlier version of our approach was ranked 5th (out of 53 entries) in the TADPOLE challenge in 2019. The current approach is ranked 2nd out of 56 entries as of August 12th, 2019.

1 Introduction

Alzheimer's disease (AD) dementia is a devastating neurodegenerative disease with a long prodromal phase and no available cure. It is widely believed that an effective treatment strategy should target individuals at risk for AD early in the disease process (Scheltens et al., 2016). Consequently, there is significant interest in predicting the longitudinal disease progression of individuals. A major difficulty is that although AD commonly presents as an amnestic syndrome, there is significant heterogeneity across individuals (Murray et al., 2011; Noh et al., 2014; Zhang et al., 2016; Risacher et al., 2017; Young et al., 2018; Sun et al., 2019). Since AD dementia is marked by beta-amyloid- and tau-mediated injuries, followed by brain atrophy and cognitive decline (Jack et al., 2010, 2013), a multimodal approach might be more effective than a single modality approach to disentangle this heterogeneity and predict longitudinal disease progression (Marinescu et al., 2018).

In this study, we proposed a machine learning algorithm to predict multimodal AD markers (e.g., ventricular volume, cognitive scores, etc) and clinical diagnosis of individual participants for every month up to six years into the future. Most previous work has focused on a "static" variant of the problem, where the goal is to predict a single timepoint (Duchesne et al., 2009; Stonnington et al., 2010; Zhang and Shen, 2012; Moradi et al., 2015; Albert et al., 2018; Ding et al., 2018) or a set of *pre-specified* timepoints in the future (regularized regression; (Wang et al., 2012; Johnson et al., 2012; McArdle et al., 2016; Wang et al., 2016)). By contrast, our goal is the longitudinal prediction of clinical diagnosis and multimodal AD markers at a potentially unlimited number of timepoints into the future[1], as defined by The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge (Marinescu et al., 2018), which arguably a more relevant and complete goal for tasks, such as prognosis and cohort selection.

One popular approach to this longitudinal prediction problem is mixed-effect regression modeling, where longitudinal trajectories of AD biomarkers are parameterized by linear or sigmoidal curves (Vemuri et al., 2009; Ito et al., 2010; Sabuncu et al., 2014; Samtani et al., 2012; Zhu and Sabuncu, 2018). However, such a modeling approach requires knowing the shapes of the biomarker trajectories a priori. Furthermore, even though the biomarker trajectories might be linear or sigmoidal when *averaged* across participants (Caroli and

[1] Although the goal is to (in principle) predict an unlimited number of time points into the future, the evaluation can only be performed using the finite number of timepoints available in the dataset.

Frisoni, 2010; Jack et al., 2010; Sabuncu et al., 2011), individual subjects might deviate significantly from the assumed parametric forms.

Consequently, it might be advantageous to not assume that the biomarker trajectories follow a specific functional form. For example, Xie and colleagues proposed an incremental regression modeling approach to predict the next timepoint based on a fixed number of input time points (Xie et al., 2016). The prediction can then be used as input to predict the next timepoint and so on indefinitely. However, the training procedure requires participants to have two timepoints, thus "wasting" data from participants with less or more than two timepoints. Therefore, state-based models that do not constrain the shapes of the biomarker trajectories or assume a fixed number of timepoints might be more suitable for this longitudinal prediction problem (e.g., discrete state hidden Markov models; Sukkar et al. 2012). Here, we considered recurrent neural networks (RNNs), which allow an individual's latent state to be represented by a vector of numbers, thus providing a richer encoding of an individual's "disease state" beyond a single integer (as in the case of discrete state hidden Markov models). In the context of medical applications, RNNs have been used to model electronic health records (Lipton et al., 2016a; Choi et al., 2016; Esteban et al., 2016; Pham et al., 2017; Rajkomar et al., 2018; Suo et al., 2018) and AD disease progression (Nguyen et al., 2018; Ghazi et al., 2019).

Most previous work on predicting AD progression ignore the issue of missing data (Stonnington et al., 2010; Sukkar et al., 2012; Lei et al., 2017; Liu et al., 2019). However, missing data is prevalent in real-world applications and arises due to study design, delay in data collection, subject attrition or mistakes in data collection. Missing data poses a major difficulty for modeling longitudinal data since most statistical models assume feature-complete data (García-Laencina et al., 2010). Many studies sidestep this issue by removing subjects or timepoints with missing data, thus potentially losing a large quantity of data. There are two main approaches for handling missing data (Schafer and Graham 2002). First, the "preprocessing" approach handles the missing data issue in a separate preprocessing step, by imputing the missing data (e.g., using the missing variable's mean or more sophisticated machine learning strategies; Azur et al., 2011; Rehfeld et al., 2011; Stekhoven and Bühlmann, 2011; White et al., 2011; Zhou et al., 2013), and then using the imputed data for subsequent modeling. Second, the "integrative" approach is to integrate the missing data issue directly into the models or training strategies, e.g., marginalizing the missing data in Bayesian approaches (Marquand et al., 2014; Wang et al., 2014; Aksman et al., 2019).

In this work, we proposed to adapt the minimalRNN model (Chen, 2017) to predict AD progression. The minimalRNN has fewer parameters than other RNN models, such as the long short-term memory (LSTM) model, so it might be less prone to overfitting. Although RNNs are usually trained using feature-complete data, we explored two “preprocessing” and one “integrative” approaches to deal with missing data. We used data from the TADPOLE competition, comprising longitudinal data from 1677 participants (Marinescu et al. 2018). An earlier version of this work was published at the International Workshop on Pattern Recognition in Neuroimaging and utilized the more complex LSTM model (Nguyen et al., 2018). Here, we extended our previous work by using a simpler RNN model, expanding our comparisons with baseline approaches and exploring how the number of input timepoints affect prediction performance. We also compared the original LSTM and current minimalRNN models using the live leaderboard on TADPOLE.

2 Methods

2.1 Problem setup

The problem setup follows that of the TADPOLE challenge (Marinescu et al. 2018). Given the multimodal AD markers and diagnostic status of a participant from one or more timepoints, we seek to predict the cognition (as measured by ADAS-Cog13; Mohs et al., 1997), ventricular volume (as measured by structural MRI) and clinical diagnosis of the participant for every month indefinitely into the future.

2.2 Data

We utilized the data provided by the TADPOLE challenge (Marinescu et al., 2018). The data consisted of 1677 subjects from the ADNI database (Jack et al., 2008). Each participant was scanned at multiple timepoints. The average number of timepoints was 7.3 ± 4.0 (Figure 1A), while the average number of years from the first timepoint to the last timepoint was 3.6 ± 2.5 (Figure 1B).

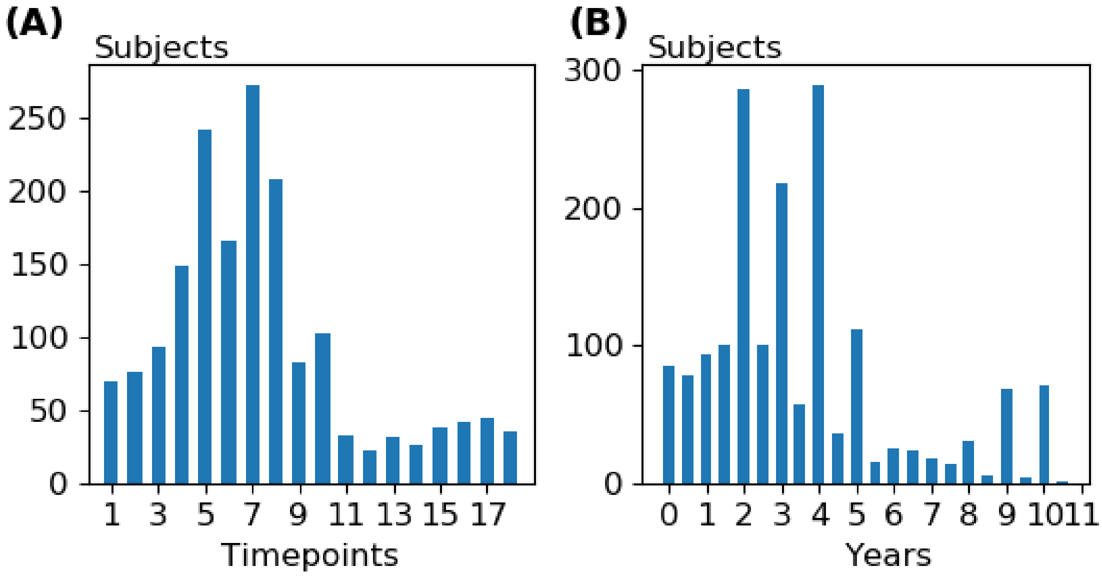


Figure 1. (A) Distribution of the number of timepoints for all subjects in the dataset. **(B)** Distribution of the number of years between the first and last timepoints for all subjects in the dataset.

For consistency, we used the same set of 23 variables recommended by the TADPOLE challenge, which included diagnosis, neuropsychological test scores, anatomical features derived from T1 magnetic resonance imaging (MRI), positron emission tomography

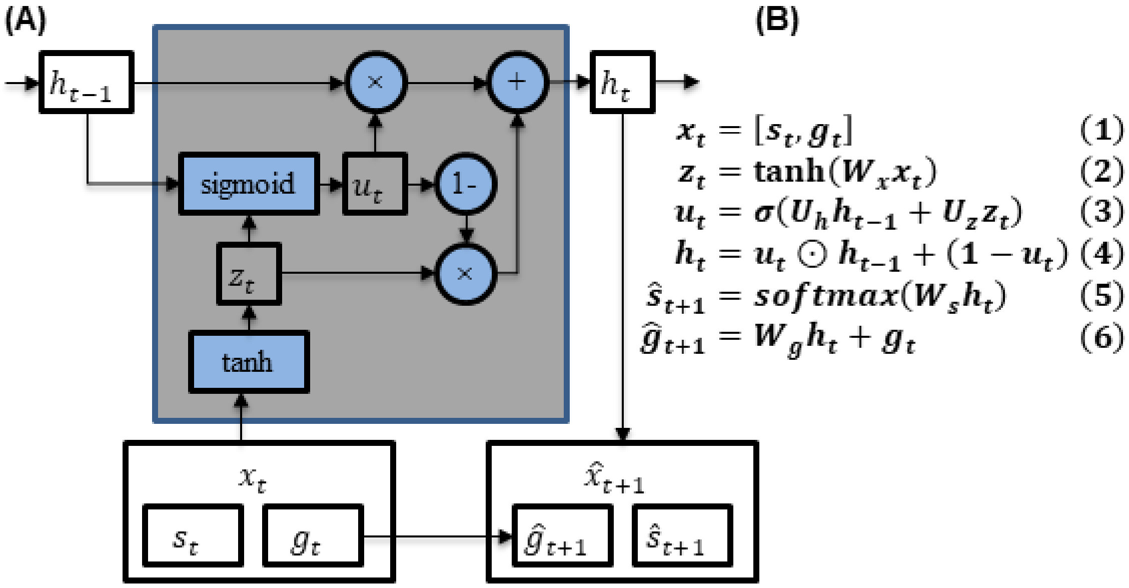
(PET) measures and CSF markers (Table 1). The diagnostic categories corresponded to normal control (NC), mild cognitive impairment (MCI) and Alzheimer's disease (AD).

	mean (± std)	% timepoints with measures
Clinical Dementia Rating Scale (SB)	$2.17 \pm 2.81 \times 10^0$	70.36 %
ADAS-Cog11	$1.13 \pm 0.86 \times 10^1$	69.95 %
ADAS-Cog13	$1.75 \pm 1.16 \times 10^1$	69.27 %
Mini-Mental State Examination (MMSE)	$2.65 \pm 0.39 \times 10^1$	70.12 %
RAVLT immediate	$3.44 \pm 1.36 \times 10^1$	69.33 %
RAVLT learning	$4.02 \pm 2.81 \times 10^0$	69.33 %
RAVLT forgetting	$4.23 \pm 2.52 \times 10^0$	69.12 %
RAVLT forgetting percent	$5.97 \pm 3.83 \times 10^1$	68.57 %
Functional Activities Questionnaire (FAQ)	$5.59 \pm 7.92 \times 10^0$	70.60 %
Montreal Cognitive Assessment (MOCA)	$2.30 \pm 0.47 \times 10^1$	38.99 %
Ventricles	$4.21 \pm 2.32 \times 10^4$	58.44 %
Hippocampus	$6.68 \pm 1.24 \times 10^3$	53.39 %
Whole brain volume	$1.01 \pm 0.11 \times 10^6$	60.35 %
Entorhinal cortical volume	$3.44 \pm 0.81 \times 10^3$	50.78 %
Fusiform cortical volume	$1.71 \pm 0.28 \times 10^4$	50.78 %
Middle temporal cortical volume	$1.92 \pm 0.31 \times 10^4$	50.78 %
Intracranial volume	$1.53 \pm 0.16 \times 10^6$	62.43 %
Florbetapir (18F-AV-45) - PET	$1.19 \pm 0.22 \times 10^0$	16.62 %
Fluorodeoxyglucose (FDG) - PET	$1.20 \pm 0.16 \times 10^0$	26.31 %
Beta-amyloid (CSF)	$1.02 \pm 0.59 \times 10^3$	18.60 %
Total tau	$2.93 \pm 1.30 \times 10^2$	18.55 %
Phosphorylated tau	$4.80 \pm 1.44 \times 10^1$	18.62 %
Diagnosis	-	69.89 %

Table 1. Set of variables together with their means, standard deviations and percentage of timepoints where the variables were actually observed. SB: Sum of boxes, ADAS: Alzheimer's Disease Assessment Scale, RAVLT: Rey Auditory Verbal Learning Test

We randomly divided the data into training, validation and test sets. The ratio of subjects in the training, validation and test sets was 18:1:1. The training set was used to train the model. The validation set was used to select the hyperparameters. The test set was used to evaluate the models' performance. For subjects in the validation and test sets, the first half of the timepoints of each subject were used to predict the second half of the timepoints of the same subject. All variables (except diagnostic category, which was categorical rather than continuous) were z-normalized. The z-normalization was performed on the training set. The mean and standard deviation from the training set was then utilized to z-normalize the validation and test sets. The random split of the data into training, validation and test sets was repeated 20 times to ensure stability of results (Kong et al., 2019; Li et al., 2019; Varoquaux, 2018). Care was taken so that the test sets were non-overlapping so that the test sets across the 20 data splits covered the entire dataset.

2.3 Proposed model



$$\boldsymbol{x_t} = [\boldsymbol{s_t}, \boldsymbol{g_t}] \quad (1)$$
$$\boldsymbol{z_t} = \tanh(\boldsymbol{W_x x_t}) \quad (2)$$
$$\boldsymbol{u_t} = \boldsymbol{\sigma}(\boldsymbol{U_h h_{t-1}} + \boldsymbol{U_z z_t}) \quad (3)$$
$$\boldsymbol{h_t} = \boldsymbol{u_t} \odot \boldsymbol{h_{t-1}} + (\boldsymbol{1} - \boldsymbol{u_t}) \quad (4)$$
$$\boldsymbol{\hat{s}_{t+1}} = \boldsymbol{softmax}(\boldsymbol{W_s h_t}) \quad (5)$$
$$\boldsymbol{\hat{g}_{t+1}} = \boldsymbol{W_g h_t} + \boldsymbol{g_t} \quad (6)$$

Figure 2. (A) MinimalRNN. **(B)** MinimalRNN update equations. $\boldsymbol{s_t}$ and $\boldsymbol{g_t}$ denote categorical (i.e., diagnosis) and continuous variables respectively (Table 1). The input $\boldsymbol{x_t}$ to each RNN cell comprised the diagnosis $\boldsymbol{s_t}$ and continuous variables $\boldsymbol{g_t}$ (Eq. 1). Note that $\boldsymbol{s_t}$ was represented using one-hot encoding. The hidden state $\boldsymbol{h_t}$ was a combination of the previous hidden state $\boldsymbol{h_{t-1}}$ and the transformed input $\boldsymbol{z_t}$ (Eq. 4). The forget gate $\boldsymbol{u_t}$ weighed the contributions of the previous hidden state $\boldsymbol{h_{t-1}}$ and current transformed input $\boldsymbol{z_t}$ toward the current hidden state $\boldsymbol{h_t}$ (Eq. 3). The model predicted the next month diagnosis $\boldsymbol{\hat{s}_{t+1}}$and continuous variables $\boldsymbol{\hat{g}_{t+1}}$ using the hidden state $\boldsymbol{h_t}$ (Eqs. 5 and 6). $\odot$ and $\boldsymbol{\sigma}$ denote element-wise product and the sigmoid function respectively.

We adapted the minimalRNN (Chen, 2017) for predicting disease progression. The model architecture and update equations are shown in Figure 1. Let $\boldsymbol{x_t}$ denote all variables observed at time t, comprising the diagnosis $\boldsymbol{s_t}$ and remaining continuous variables $\boldsymbol{g_t}$ (Eq. 1 in Figure 2B). Here, diagnosis was represented using one-hot encoding. In other words, diagnosis was represented as a vector of length three. More specifically, if the first entry was one, then the participant was a normal control. If the second entry was one, then the participant was mild cognitively impaired. If the third entry was one, then the participant had

AD dementia. For now, we assume that all variables were observed at all timepoints; the missing data issue will be addressed in Sections 2.4.

At each timepoint, the transformed input $\boldsymbol{z_t}$ (Eq. 2 in Figure 2) and the previous hidden state $\boldsymbol{h_{t-1}}$ were used to update the hidden state $\boldsymbol{h_t}$ (Eqs. 3 and 4 in Figure 2B). The hidden state can be interpreted as integrating all information about the subject up until that timepoint. The hidden state $\boldsymbol{h_t}$ was then used to predict the observations at the next timepoint $\boldsymbol{x_{t+1}}$ (Eqs. 5 and 6 in Figure 1B).

In the ADNI database, data were collected at a minimum interval of 6 months. However, in practice, data might be collected at an unscheduled time (e.g., month 8 instead of month 6). Consequently, the duration between timepoints t and $t+1$ in the RNN was set to be 1 month.

2.3.1 Training with no missing data

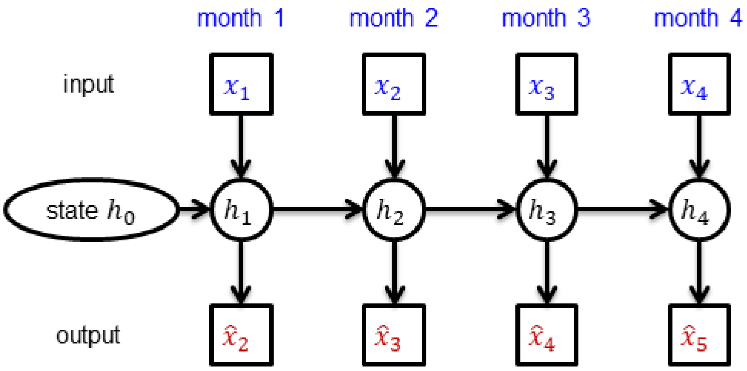


Figure 3. The minimalRNN was trained to predict the next observation given the current observation (e.g., predicting $\boldsymbol{\hat{x}_2}$ given $\boldsymbol{x_1}$). The errors between the actual observations (e.g., $\boldsymbol{x_2}$) and predictions (e.g., $\boldsymbol{\hat{x}_2}$) were used to update the model parameters. The hidden state $\boldsymbol{h_t}$ encoded information about the subject up until time t.

The RNN training is illustrated in Figure 2. The RNN was trained to predict the next observation ($\boldsymbol{x_t}$) given the previous observations ($\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_{t-1}}$). The errors between the predicted outputs (e.g. $\boldsymbol{\hat{x}_2}$) and the ground truth outputs (e.g. $\boldsymbol{x_2}$) were used to update the model parameters. The error (or loss L) was defined as follows:

$$L = \sum_{t>1} (\text{CrossEntropy}(\boldsymbol{s_t}, \boldsymbol{\hat{s}_t}) + \text{MAE}(\boldsymbol{g_t}, \boldsymbol{\hat{g}_t})) \quad (7)$$

$$\text{CrossEntropy}(\boldsymbol{s_t}, \boldsymbol{\hat{s}_t}) = -\sum_{j=1}^{3} \boldsymbol{s_t^j} \log \boldsymbol{\hat{s}_t^j} \quad (8)$$

$$\text{MAE}(\boldsymbol{g_t}, \boldsymbol{\hat{g}_t}) = |\boldsymbol{g_t} - \boldsymbol{\hat{g}_t}| \quad (9)$$

The value of $\boldsymbol{h_0}$ was set to be $\boldsymbol{0}$. During training, gradients of loss L with respect to the model parameters were back-propagated to update the RNN parameters. The RNN was trained using Adam (Kingma and Ba, 2015). The HORD algorithm (Regis and Shoemaker 2013; Eriksson, Bindel, and Shoemaker 2015; Ilievski et al. 2017) was utilized to find the best hyperparameters by maximizing model performance on the validation set. We note that this optimization was performed independently for each training/validation/test split of the dataset. The hyperparameter search space is shown in Table 2.

Hyper-parameter	Range
Input dropout rate	0.0 – 0.5
Recurrent dropout rate	0.0 – 0.5
L2 weight regularization	10^{-7} – 10^{-5}
Learning rate	10^{-5} – 10^{-2}
Number of hidden layers	1 – 3
Size of hidden state	128 – 512

Table 2. Hyperparameter search space of the MinimalRNN estimated from the validation sets using HORD.

2.3.2 Prediction with no missing data

Figure 3 illustrates how the RNN was used to predict AD progression in an example subject (from the validation or test set). Given observations for months 1, 2 and 3, the goal of the model was to predict observations in future months. From month 4 onwards, the model predictions ($\hat{\boldsymbol{x}}_4$ and $\hat{\boldsymbol{x}}_5$) were fed in as inputs to the RNN (for months 5 and 6 respectively) to make further predictions (dashed lines in Figure 4).

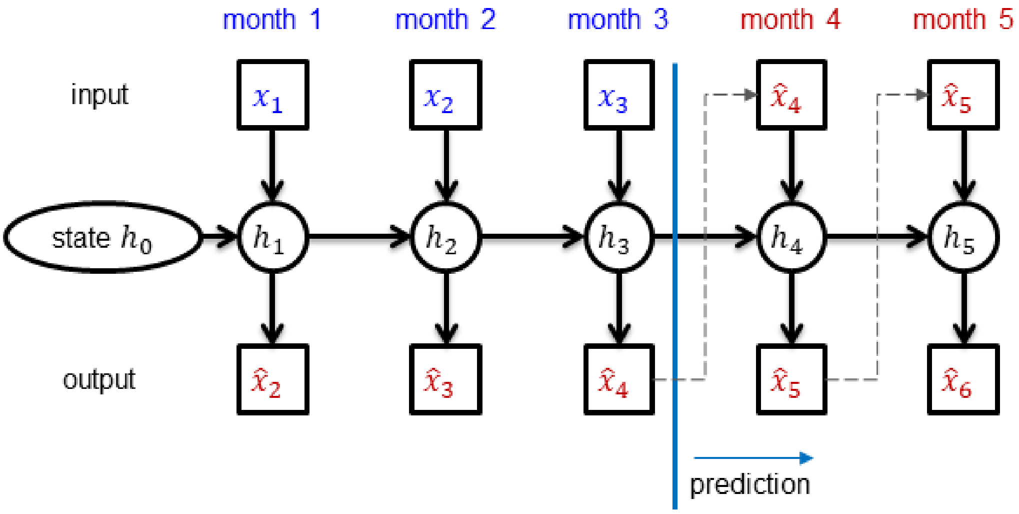


Figure 4. Predicting future timepoints ($\hat{\boldsymbol{x}}_4$, $\hat{\boldsymbol{x}}_5$, $\hat{\boldsymbol{x}}_6$, etc) given three initial timepoints ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$). Prediction started at month 4. Since there were no observed data at timepoints 4 and 5, the predictions ($\hat{\boldsymbol{x}}_4$ and $\hat{\boldsymbol{x}}_5$) were used as inputs (at timepoints 5 and 6 respectively) to predict further into the future.

2.4 Missing data

As seen in Table 1, there were a lot of missing data in ADNI. This was exacerbated by the fact that data were collected at a minimum interval of 6 months, while the sampling period in the RNN was set to be one month (to handle off-schedule data collection). During training, the loss function was evaluated only at timepoints with available observations. Similarly, when evaluating model performance (Section 2.6), only available observations were utilized.

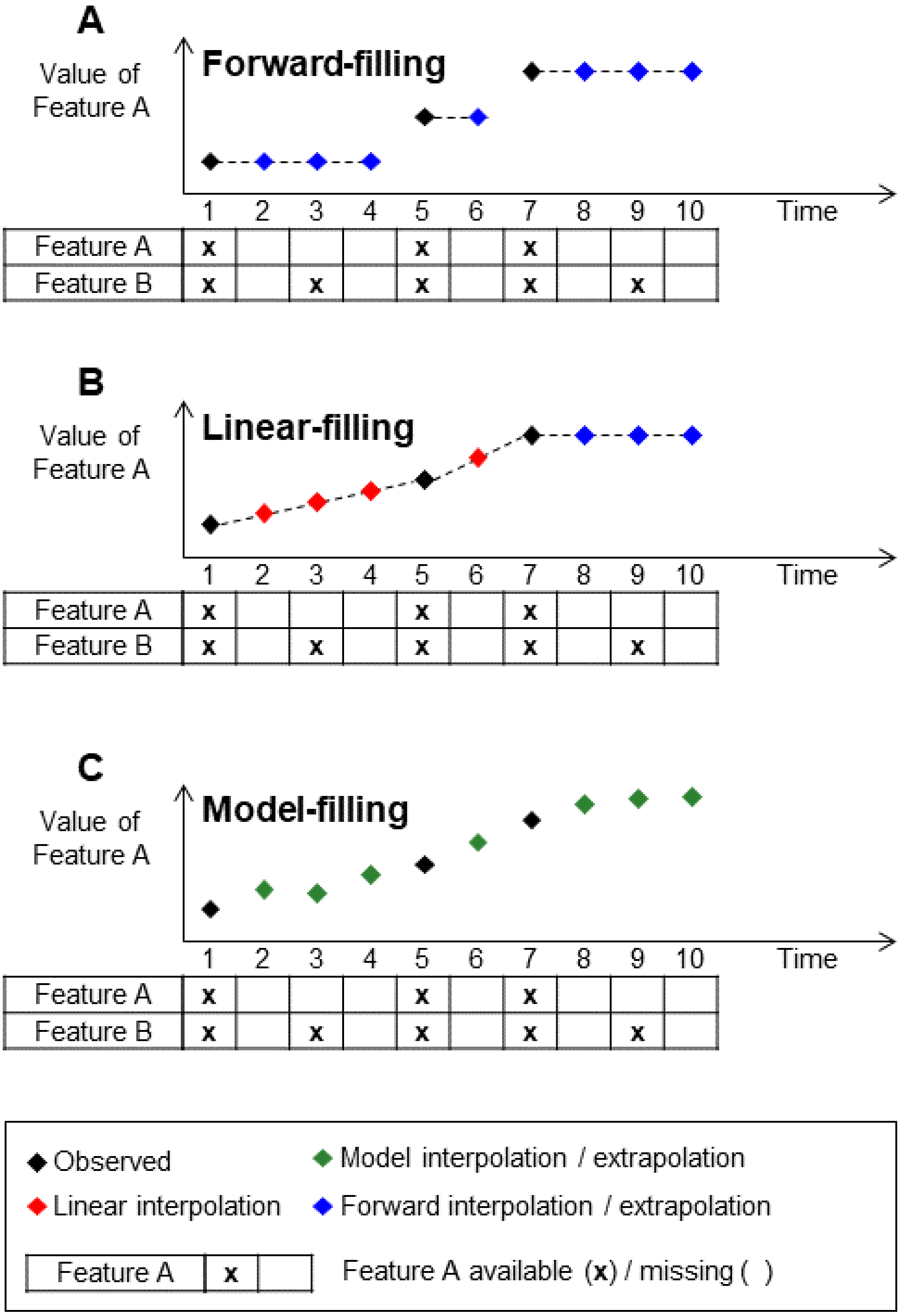
The missing data also posed a problem for the RNN update equations (Figure 1B), which assumed all variables were observed. Here, we explored two "preprocessing" strategies (Sections 2.4.1 & 2.4.2) and one "integrative" strategy (Section 2.4.3) to handle the missing values. As explained in the introduction, "preprocessing" strategies impute the missing data in a separate preprocessing. The imputed data is then used for subsequent modeling. On the other hand, "integrative" strategies incorporate the missing data issue directly into the model or training strategies.

2.4.1 Forward filling

Forward filling involved imputing the data using the last timepoint with available data (Che et al., 2018; Lipton et al., 2016b). Figure 5A illustrates an example of how forward-filling in time was used to fill in missing input data. In this example, there were two input variables A and B. The values of feature A at time t = 2, 3 and 4 were filled using the last observed value of feature A (at time t = 1). Similarly, the values at t = 7, 8 of feature A were filled using value at t = 6 when it was last observed. If data was missing at the first timepoint, the mean value across all timepoints of all training subjects was used for the imputation.

2.4.2 Linear filling

The previous strategy utilized information from previous timepoints for imputation. One could imagine that it might be helpful to use previous and future timepoints for imputation. The linear filling strategy performed linear interpolation between the previous timepoint and the next time point with available data (Junninen et al., 2004). Figure 5B shows an example of linear interpolation. Values of feature A at time t = 2, 3, 4, 6 were filled in using linear interpolation. However, linear-filling did not work for months 8, 9 and 10 because there was no future observed data for linear interpolation, so forward-filling was utilized for those timepoints. Like forward filling, if data was missing at the first timepoint, the mean value across all timepoints of all training subjects was used for the imputation.



	1	2	3	4	5	6	7	8	9	10
Feature A	x				x		x			
Feature B	x		x		x		x		x	

Feature A	x	

Figure 5. Different strategies to impute missing data. (A) Forward-filling imputed missing values using the last observed value. (B) Linear-filling imputed missing values using linear interpolation between previous observed and next observed values. Notice that linear-filling did not work for months 8, 9 and 10 because there was no future observed data for linear interpolation, so forward filling was utilized for those timepoints. (C) Model-filling imputed missing values using model predictions.

2.4.3 Model filling

We also considered a novel model filling strategy of filling in missing data. As seen in Section 2.3.2 (Figure 5), the prediction of the RNN could be used as inputs for the next timepoint. The same approach can be used for filling in missing data.

Figure 5B shows an example of how the RNN was used to fill in missing data. At time t = 2 to 6, the values of feature A were filled in using predictions from the RNN. The RNN could also be used to extrapolate features that "terminate early" (e.g., time t = 8 and 9).

A theoretical benefit of modeling filling was that the full sets of features were utilized for the imputation. For example, both features A and B at time t = 1 were used by the RNN to predict both input features at time t = 2 (Figure 5B). This was in contrast to forward or linear filling, which would utilize only feature A (or B) to impute feature A (or B).

Like forward filling, if data was missing at the first timepoint, the mean value across all timepoints of all training subjects was used for the imputation.

2.5 Baselines

We considered three baselines: constant prediction, support vector machine/regression (SVM/SVR), and linear state-space (LSS) model.

2.5.1 Constant prediction

The constant prediction algorithm simply predicted all future values to be the same as the last observed values. The algorithm did not need any training. While this might seem like an overly simplistic algorithm, we will see that the constant prediction algorithm is quite competitive for near term prediction.

2.5.2 SVM/SVR

As explained in the introduction, most previous studies have focused on a "static" variant of the problem, where the goal is to predict a single timepoint or a set of pre-specified timepoints in the future. Here, we will consider such a baseline by using SVM to predict clinical diagnosis (which was categorical) and SVR to predict ADAS-Cog13 and ventricular volume (which were continuous). The models were implemented using scikit-learn (Pedregosa et al., 2011).

Overall, we considered four SVM/SVR baselines corresponding to using 1 to 4 input timepoints (spaced 6 months apart) to predict the future. The 6-month interval was chosen because the ADNI data was collected roughly every 6 months. As can be seen in Section 3.1, the best results were obtained with 2 or 3 input timepoints, so we did not explore more than 4 input timepoints.

For each SVM/SVR baseline, we trained separate SVM/SVR to predict 6, 12, 18, …, 60 months into the future. 60 months were the maximum because of insufficient data to train SVM/SVR to predict further into the future (Figure 1B). Therefore, in total, for each of the four SVM/SVR baselines (using 1, 2, 3 or 4 input timepoints), we trained 10 separate SVM to predict clinical diagnosis, 10 SVR to predict ADAS-Cog13 and 10 SVR to predict ventricular volume.

The linear filling strategy (Figure 5B) was used to handle missing data. Because prediction performance was evaluated at every month in the future, prediction at intermediate months (e.g., months 1 to 5, 7 to 11, etc) were linearly interpolated. Prediction from month 61 onwards utilized forward filling based on the prediction at month 60.

We used the same 20 training/validation/test data splits as the RNN. For each data split, the SVM/SVR was trained on the training set and the hyperparameters were selected using the validation set using HORD. The models were then evaluated in the test set.

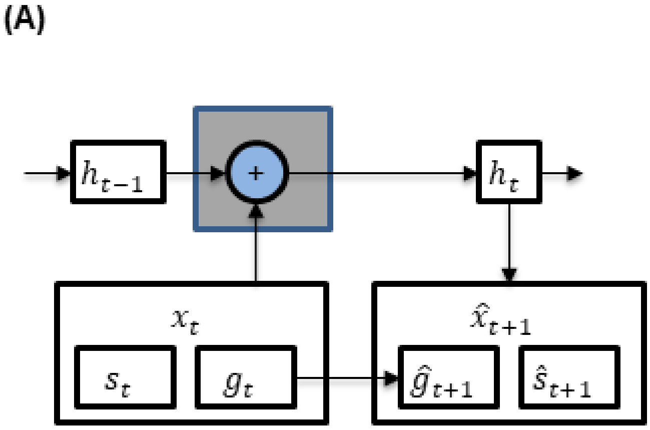
	SVM	**SVR**
Kernel	Linear or RBF	
Epsilon	NA	$10^{-3} – 10^{-0}$
Penalty	$10^{-3} – 10^{3}$	
Gamma	$10^{-3} – 10^{3}$	

Table 3. Hyperparameter search space of the SVM/SVR estimated from the validation sets using HORD.

One tricky issue arose when a test subject had insufficient input timepoints for a particular SVM/SVR baseline. For example, the 4-timepoint SVM/SVR baseline required 4 input timepoints in order to predict future timepoints. In this scenario, if a test subject only had 2 input timepoints, then the 2-timepoint SVM/SVR was utilized for this subject even though we were considering the 4-timepoint SVM/SVR baseline. We utilized this strategy (instead of discarding the test subject) in order to ensure the test sets were exactly the same across all algorithms.

2.5.3 Linear state space (LSS) model

We considered a linear state space (LSS) baseline by linearizing the minimalRNN model (Figure 6). Other than the update equations (Figure 6), all other aspects of training and prediction were kept the same. For example, the LSS models utilized the same data imputation strategies (Section 2.4) and were trained with the same cost function using Adam. We used the same 20 training/validation/test data splits as the RNN. HORD was used to find the best hyperparameters by maximizing the performance in the validation sets. The search ranges of the hyperparameters were the same as before (Table 2).



(B)

$$\boldsymbol{x}_t = [\boldsymbol{s}_t, \boldsymbol{g}_t] \qquad (10)$$

$$\boldsymbol{h}_t = \boldsymbol{W}_h \boldsymbol{h}_{t-1} + \boldsymbol{W}_x \boldsymbol{x}_t \qquad (11)$$

$$\hat{\boldsymbol{s}}_{t+1} = \boldsymbol{softmax}(\boldsymbol{W}_s \boldsymbol{h}_t) \qquad (12)$$

$$\hat{\boldsymbol{g}}_{t+1} = \boldsymbol{W}_g \boldsymbol{h}_t + \boldsymbol{g}_t \qquad (13)$$

Figure 6. (A) Linear state space (LSS) model. Observe the gray cell is much simpler than the minimalRNN **(B)** LSS update equations. $\boldsymbol{s}_t$ and $\boldsymbol{g}_t$ denote categorical (i.e., diagnosis) and continuous variables respectively (Table 1). The input $\boldsymbol{x}_t$ to each LSS cell comprised the diagnosis $\boldsymbol{s}_t$ and continuous variables $\boldsymbol{g}_t$ (Eq. 10). Like before, $\boldsymbol{s}_t$ was represented using one-hot encoding. The hidden state $\boldsymbol{h}_t$ was a combination of the previous hidden state $\boldsymbol{h}_{t-1}$ and the input $\boldsymbol{x}_t$ (Eq. 11). The model predicted the next month diagnosis $\hat{\boldsymbol{s}}_{t+1}$ and continuous variables $\hat{\boldsymbol{g}}_{t+1}$ using the hidden state $\boldsymbol{h}_t$ (Eqs. 12 and 13).

2.6 Performance evaluation

Following the TADPOLE competition, diagnosis classification accuracy was evaluated using the multiclass area under the operating curve (mAUC; Hand and Till, 2001) and balanced class accuracy (BCA) metrics. For both mAUC and BCA metrics, higher values indicate better performance. ADAS-Cog13 and ventricles prediction accuracy was evaluated using mean absolute error (MAE). Lower MAE indicates better performance. The final performance for each model was computed by averaging the results across the 20 test sets.

The resampled t-test (Bouckaert and Frank, 2004) was used to evaluate differences in performance between models.

2.7 Impact of the number of input timepoints on prediction accuracy

For the RNNs to be useful in the clinical settings, they should ideally be able to perform well with as little input timepoints as possible. Therefore, we applied the best model (Section 2.6) to the test subjects using 1, 2, 3 or 4 input timepoints (Figure 9). Test subjects with less than 4 input timepoints were discarded, so that the same test subjects were evaluated across the four conditions (i.e., 1, 2, 3 or 4 input timepoints). However, this meant that the results from this analysis were not comparable to the previous sections (since the test subjects were not exactly the same).

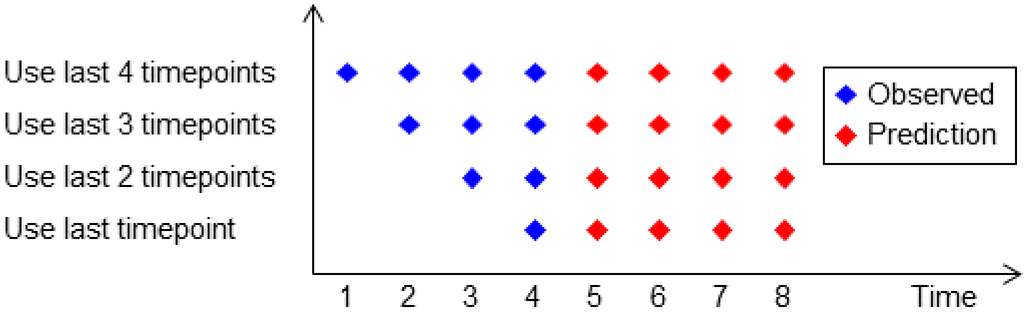


Figure 7. Prediction performance as a function of the number of input timepoints in the test subjects.

2.8 TADPOLE live leaderboard

The TADPOLE challenge involves the prediction of ADAS-Cog13, ventricular volume and clinical diagnosis of 219 ADNI participants for every month up to five years into the future. We note that these 219 participants were a subset of the 1677 subjects used in this study. However, the future timepoints used to evaluate performance on the live leaderboard (https://tadpole.grand-challenge.org/D4_Leaderboard/) were not part of the data utilized in this study. Here, we utilized the entire dataset (1677 participants) to tune a set of hyperparameters (using HORD) that maximized performance either (1) one year into the future or (2) all years into the future. We then submitted the predictions of the 219 participants to the TADPOLE leaderboard.

2.9 Data and code availability

The code used in this paper can be found at https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/Ng uyen2020_RNNAD. This study utilized data from the publicly available ADNI database (http://adni.loni.usc.edu/data-samples/access-data/). The particular set of participants and features we used is available at the TADPOLE website (https://tadpole.grand-challenge.org/).

3 Results

3.1 Overall performance

Figure 8 illustrates the test performance of minimalRNN and three baselines (SVM/SVR, constant prediction and LSS). For clarity, we only showed RNN with mixed filling (RNN–MF), LSS with mixed filling (LSS–MF) and SVM/SVR using one input timepoint because they yielded the best results within their model classes. Table 4 shows the test performance of all models (RNN, SVM/SVR, constant prediction and LSS) across all three missing data strategies.

We performed statistical tests comparing the three RNN variants (RNN–FF, RNN–LF and RNN–MF) with all other baseline approaches (LSS, constant prediction, SVM/SVR). Multiple comparisons were corrected with a false discovery rate (FDR) of $q < 0.05$. In the case of clinical diagnosis prediction, RNN–MF performed the best and was statistically better than all baseline approaches (LSS, constant prediction, SVM/SVR). In the case of ADAS-Cog13 and ventricular volume, RNN-MF also performed the best and was statistically better than all baseline approaches, except LSS with model filling (LSS–MF; $p = 0.59$).

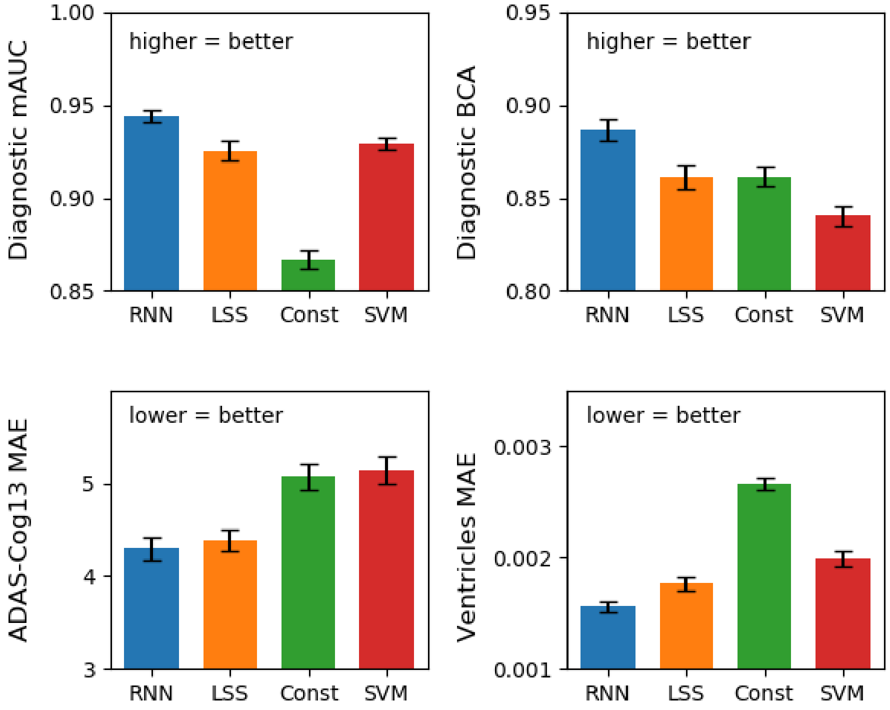


Figure 8. Performance of the best models from each model class averaged across 20 test sets. Error bars show standard error across test sets. For clinical diagnosis, higher mAUC and BCA values indicate better performance. For ADAS-Cog13 and Ventricles, lower MAE indicates better performance. The RNN model corresponded to RNN–MF in Table 4. The

SVM model corresponded to SVM/SVR (= 1tp) model in Table 4. The LSS model corresponded to LSS–MF in Table 4. RNN performed the best. See Figure S1 for all models.

	mAUC (more=better)	BCA (more=better)	ADAS-Cog13 (less=better)	Ventricles (less=better)
RNN–FF	0.923 ± 0.019	0.867 ± 0.023	5.03 ± 0.62	0.00247 ± 0.00036
RNN–LF	0.910 ± 0.031	0.858 ± 0.028	5.42 ± 0.94	0.00193 ± 0.00029
RNN–MF	**0.944** ± 0.014	**0.887** ± 0.024	**4.30** ± 0.53	**0.00156** ± 0.00022
LSS–FF	0.928 ± 0.020 (p = 0.018)	0.864 ± 0.024 (p = 0.001)	4.95 ± 0.57 (p = 0.003)	0.00216 ± 0.00031 (p = 5.6×10^{-7})
LSS–LF	0.908 ± 0.032 (p = 0.005)	0.857 ± 0.037 (p = 0.042)	6.36 ± 0.82 (p = 3.2×10^{-7})	0.00175 ± 0.00023 (p = 0.061)
LSS–MF	0.926 ± 0.025 (p = 0.004)	0.861 ± 0.029 (p = 0.001)	4.38 ± 0.49 (p = 0.590)	0.00177 ± 0.00028 (p = 0.044)
Constant	0.867 ± 0.022 (p = 3.2×10^{-9})	0.861 ± 0.023 (p = 2.0×10^{-4})	5.07 ± 0.61 (p = 3.3×10^{-4})	0.00266 ± 0.00027 (p = 5.9×10^{-12})
SVM/SVR (= 1tp)	0.929 ± 0.013 (p = 0.011)	0.841 ± 0.023 (p = 2.5×10^{-7})	5.14 ± 0.62 (p = 1.8×10^{-4})	0.00199 ± 0.00031 (p = 7.3×10^{-5})
SVM/SVR (≤ 2tp)	0.926 ± 0.013 (p = 0.002)	0.836 ± 0.026 (p = 2.8×10^{-6})	5.23 ± 0.63 (p = 1.1×10^{-4})	0.00230 ± 0.00037 (p = 2.7×10^{-7})
SVM/SVR (≤ 3tp)	0.923 ± 0.013 (p = 0.001)	0.830 ± 0.025 (p = 2.6×10^{-7})	5.53 ± 0.55 (p = 4.5×10^{-7})	0.00261 ± 0.00037 (p = 5.9×10^{-7})
SVM/SVR (≤ 4tp)	0.919 ± 0.012 (p = 2.2×10^{-5})	0.832 ± 0.019 (p = 4.1×10^{-7})	5.68 ± 0.58 (p = 9.4×10^{-7})	0.00269 ± 0.00035 (p = 1.2×10^{-9})

Table 4. Prediction performance averaged across 20 test sets. For clinical diagnosis, higher mAUC and BCA values indicate better performance. For ADAS-Cog13 and Ventricles, lower MAE indicates better performance. FF indicates forward filling. LF indicates linear filling. MF indicates model filling. SVM/SVR (= 1tp) utilized one input timepoint. SVM/SVR (≤ 2tp) utilized at most 2 input timepoints (see Section 2.5.2 for details) and so on. The best result for each performance metric was **bolded**. RNN–MF was numerically the best across all metrics. Gray font indicates that the performance was not statistically better than the best performance (in bold) after correcting for multiple comparisons using a false discovery rate (FDR) of q < 0.05.

For both RNN and LSS, mixed filling performed better than forward filling and linear filling, especially when predicting ADAS-Cog13 and ventricular volume (Table 4). Interestingly, more input timepoints do not necessarily lead to better prediction in the case of SVM/SVR. In fact, the SVM/SVR model using one timepoint was numerically better than SVM/SVR models using more timepoints, although the differences were small.

Figure 9 shows the breakdown of the prediction performance from Figure 8 in yearly interval up to 6 years into the future. Not surprisingly, the performance of all algorithms became worse for predictions further into the future. The constant baseline was very competitive against the other models for the first year, but performance for subsequent years dropped very quickly. The RNN model was comparable or numerically better than all baseline approaches across all the years.

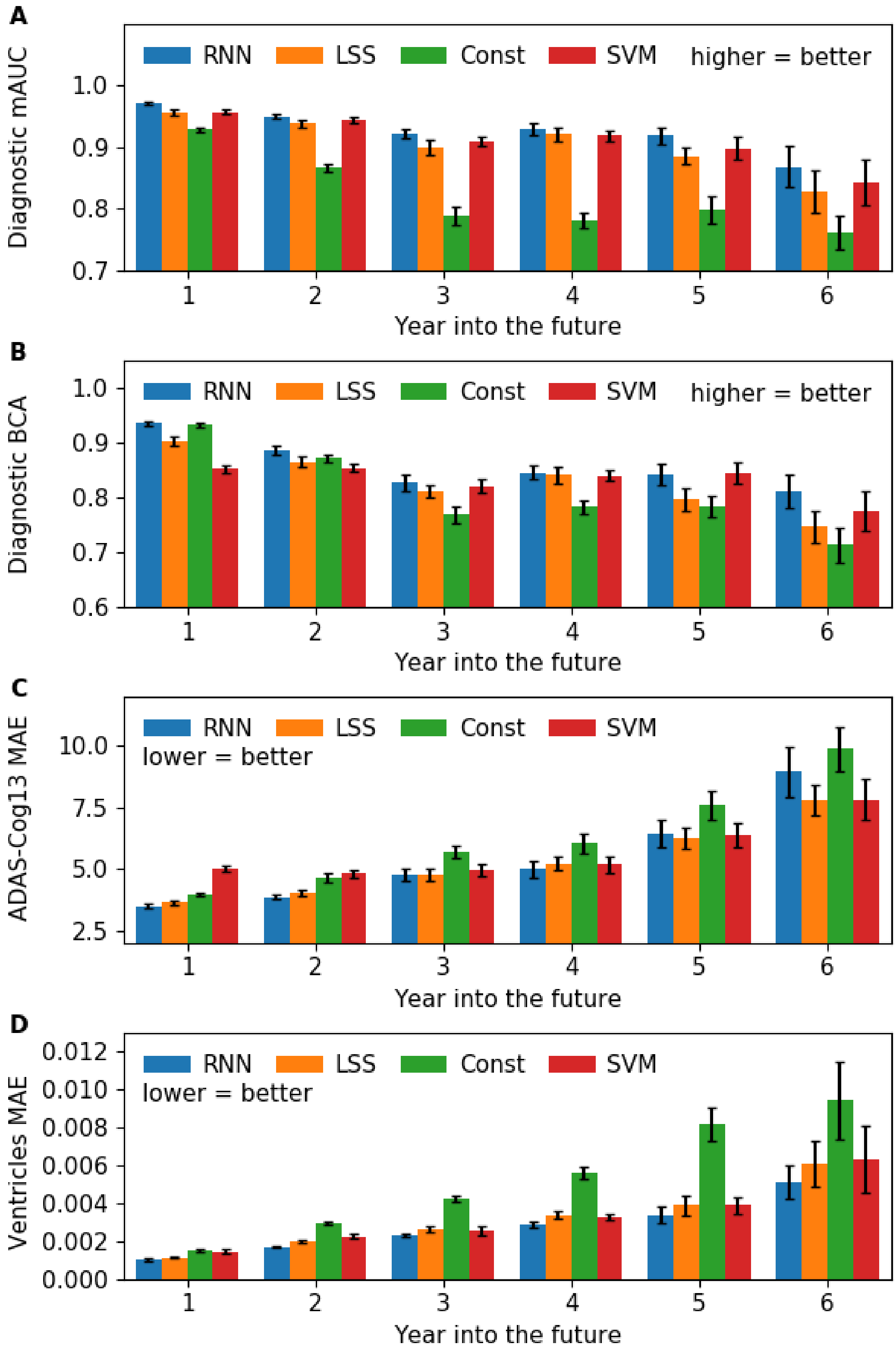


Figure 9. Prediction performance from Figure 8 broken down in yearly interval up to 6 years into the future. All algorithms became worse further into the future. RNN was comparable or numerically better across all years. See Figure S2 for all models.

3.2 RNNs using one and four input timepoints in test subjects achieve comparable performance

Given that the MinimalRNN with model filling (RNN–MF) performed the best (Table 4), we further explored how well the trained RNN–MF model would perform on test subjects with different number of input timepoints.

Figure 12 shows the performance of RNN-MF averaged across 20 test sets using different number of input timepoints. The exact numerical values are reported in Table 5. RNNs using 2 to 4 input timepoints achieved similar performance across all metrics. RNN using 1 input timepoint had numerically worse results, especially for ventricular volume. However, there was no statistical difference between using 1 input timepoint and 4 input timepoints even in the case of ventricular volume (p = 0.20).

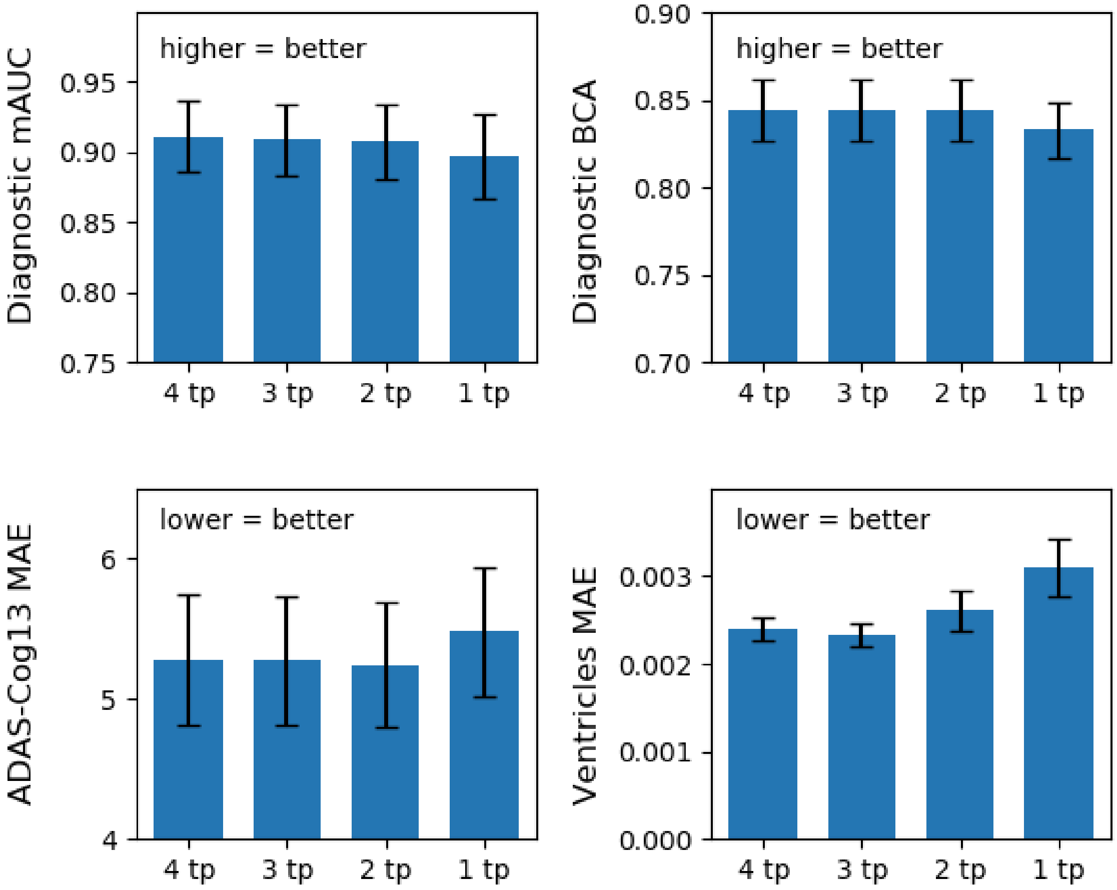


Figure 10. Test performance of RNN model with model filling strategy (RNN-MF) using different numbers of input timepoints (after training with all timepoints). Results were averaged across 20 test sets. Even though the RNN model using 1 input timepoint yielded numerically worse results, the differences were not significant (see Table 5).

	mAUC (more=better)	BCA (more=better)	ADAS-Cog13 (less=better)	Ventricles (less=better)s
4 timepoints	0.911 ± 0.076	0.844 ± 0.053	5.28 ± 1.41	0.00240 ± 0.00040
3 timepoints	0.909 ± 0.076 (p = 0.68)	0.844 ± 0.052 (p = 0.88)	5.28 ± 1.38 (p = 0.99)	0.00232 ± 0.00038 (p = 0.22)
2 timepoints	0.908 ± 0.080 (p = 0.57)	0.844 ± 0.053 (p = 0.84)	5.24 ± 1.35 (p = 0.89)	0.00260 ± 0.00067 (p = 0.50)
1 timepoint	0.897 ± 0.091 (p = 0.27)	0.833 ± 0.048 (p = 0.18)	5.48 ± 1.37 (p = 0.53)	0.00309 ± 0.00098 (p = 0.20)

Table 5. Test performance of RNN model with model filling strategy (RNN-MF) using different numbers of input timepoints (after training with all timepoints). Results were averaged across 20 test sets. Statistical tests were performed to test for differences between using 4 timepoints versus less timepoints. The gray font indicates that there was no statistical difference that survived FDR of q < 0.05.

3.3 TADPOLE live leaderboard

The original LSTM model (Nguyen et al., 2018) was ranked 5th (out of 53 entries) in the TADPOLE grand challenge in July 2019 (entry "CBIL" in https://tadpole.grand-challenge.org/Results/). Our current minimalRNN models were ranked 2nd and 3rd (out of 56 entries) in the leaderboard as of Aug 12th, 2019 (entries ("CBIL-MinMFa" and "CBIL-MinMF1"; https://tadpole.grand-challenge.org/D4_Leaderboard/). Interestingly, the model obtained from hyperparameters tuned to predict all years into the future ("CBIL-MinMFa") performed better than the model obtained from hyperparameters tuned to predict one year into the future ("CBIL-MinMF1"), even though the leaderboard currently utilized about one year of future data for prediction.

.

4 Discussion

In this work, we adapted a minimalRNN model for predicting longitudinal progression in AD dementia. Our approach compared favorably with baseline algorithms, such as SVM/SVR and LSS models. However, we note that there was no statistical difference between the minimalRNN and LSS for predicting ADAS-Cog13 and ventricular volume even though other studies suggested benefits of modeling non-linear interactions between features (Popescu et al., 2019).

As can be seen when setting up the SVM/SVR baseline models (Section 2.5.2), there are a lot of edge cases to consider in order to adapt a “static” prediction algorithm (e.g., SVM/SVR) to the more “dynamic” longitudinal prediction problem we considered here. For example, data is inevitably wasted because static approaches generally assume that participants have the same number of input timepoints. Therefore, for the SVM/SVR models using 4 input timepoints, we ended up with only 1454 participants out of the original 1677 participants. This might explain why the SVM/SVR model using 1 input timepoint compared favorably with the SVM/SVR model using 4 input timepoints (Table 4). Furthermore, we had to build multiple separate SVM/SVR models to predict at a fixed number of future timepoints, and performed interpolation at intermediate timepoints. By contrast, state-based models (e.g., minimalRNN or LSS) are more elegant in the sense that they handled participants with different number of timepoints and can in principle predict unlimited number of timepoints into the future.

Even though the ADNI dataset comprised participants with multiple timepoints, for the algorithm to be clinically useful, it has to be successful at dealing with missing data and participants with only one input timepoint. We found that the “integrative” approach of using the model to fill in the missing data (i.e., model filling) compared favorably with “preprocessing” approaches, such as forward filling or linear filling. However, it is possible that more sophisticated “preprocessing” approaches, such as matrix factorization (Mazumder et al., 2010; Nie et al., 2017; Thung et al., 2016) or wavelet interpolation (Mondal and Percival, 2010), might yield better results. We note that our model filling approach can also be considered as a form of matrix completion since the RNN (or LSS) was trained to minimize the predictive loss, which is equivalent to maximizing the likelihood of the training data. However, matrix completion usually assumes that the training data can be represented as a matrix that can be factorized into low-ranked or other specially-structured matrices. On the other hand, our method assumes temporal dependencies between rows in the data matrix (where each row is a timepoint).

Our best model (minimalRNN with model filling) had similar performance when using only 1 input timepoint instead of 4 input timepoints, suggesting that our approach might work well with just cross-sectional data (after training using longitudinal data). However, we might have simply lacked the statistical power to distinguish among the different conditions because of the smaller number of subjects in this experiment (see Section 2.7). Overall, there was no noticeable difference among using 2, 3 or 4 input timepoints, while the performance using 1 input timepoint appeared worse, but the difference was not statistically significant (Figure 10).

Although our approach compared favorably with the baseline algorithms, we note that any effective AD dementia treatment probably has to begin early in the disease process, potentially at least a decade before the emergence of behavioral symptoms. However, even in the case of our best model (minimalRNN with model filling), prediction performance of clinical diagnosis dropped from a BCA of 0.935 in year 1 to a BCA of 0.810 in year 6, while ventricular volume MAE increased from 0.00104 in year 1 to 0.00511 in year 6. Thus, significant improvement is needed for clinical utility.

One possible future direction is to investigate new features, e.g., those derived from diffusion MRI or arterial spin labeling. Previous studies have also suggested that different atrophy patterns (beyond the temporal lobe) might influence cognitive decline early in the disease process (Noh et al., 2014; Byun et al., 2015; Ferreira et al., 2017; Zhang et al., 2016; Risacher et al., 2017; Sun et al., 2019), so the atrophy features considered in this study (Table 1) might not be optimal. Although the new features may be correlated with currently used features, the new features might still provide complementary information when modeling AD progression (Popescu et al., 2019).

As mentioned in the introduction, an earlier version of our algorithm was ranked 5th out of 50 entries in the TADPOLE competition. Our current model was ranked 2nd out of 56 entries on the TADPOLE live leaderboard as of Aug 12th, 2019. Interestingly, the top team considered additional handcrafted features, which might have contributed to its success. Furthermore, the top team utilized a non-deep-learning algorithm XGboost (Chen and Guestrin, 2016), which might be consistent with recent work suggesting that for certain neuroimaging applications, non-deep-learning approaches might be highly competitive (He et al., 2019)

5 Conclusion

Using 1677 participants from the ADNI database, we showed that the minimalRNN model was better than other baseline algorithms for the longitudinal prediction of multimodal AD biomarkers and clinical diagnosis of participants up to 6 years into the future. We explored three different strategies to handle the missing data issue prevalent in longitudinal data. We found that the RNN model can itself be used to fill in the missing data, thus providing an integrative strategy to handle the missing data issue. Furthermore, we also found that after training with longitudinal data, the trained RNN model can perform reasonably well using one input timepoint, suggesting the approach might also work for cross-sectional data.

Acknowledgment

This work was supported by Singapore MOE Tier 2 (MOE2014-T2-2-016), NUS Strategic Research (DPRT/944/09/14), NUS SOM Aspiration Fund (R185000271720), Singapore NMRC (CBRG/0088/2015), NUS YIA and the Singapore National Research Foundation (NRF) Fellowship (Class of 2017). Our research also utilized resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896 and instruments supported by 1S10RR023401, 1S10RR019307, and 1S10RR023043 from the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital. Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). The Titan Xp used for this research was donated by the NVIDIA Corporation. 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 DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing 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 coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Aksman, L.M., Scelsi, M.A., Marquand, A.F., Alexander, D.C., Ourselin, S., Altmann, A., for ADNI, 2019. Modeling longitudinal imaging biomarkers with parametric Bayesian multi-task learning. Hum. Brain Mapp. https://doi.org/10.1002/hbm.24682

Albert, M., Zhu, Y., Moghekar, A., Mori, S., Miller, M.I., Soldan, A., Pettigrew, C., Selnes, O., Li, S., Wang, M.-C., 2018. Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years. Brain J. Neurol. 141, 877–887. https://doi.org/10.1093/brain/awx365

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple Imputation by Chained Equations: What is it and how does it work? Int. J. Methods Psychiatr. Res. 20, 40–49. https://doi.org/10.1002/mpr.329

Bouckaert, R.R., Frank, E., 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms, in: Dai, H., Srikant, R., Zhang, C. (Eds.), Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 3–12.

Byun, M.S., Kim, S.E., Park, J., Yi, D., Choe, Y.M., Sohn, B.K., Choi, H.J., Baek, H., Han, J.Y., Woo, J.I., Lee, D.Y., Initiative, A.D.N., 2015. Heterogeneity of Regional Brain Atrophy Patterns Associated with Distinct Progression Rates in Alzheimer's Disease. PLOS ONE 10, e0142756. https://doi.org/10.1371/journal.pone.0142756

Caroli, A., Frisoni, G.B., 2010. The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. Neurobiol. Aging 31, 1263–1274. https://doi.org/10.1016/j.neurobiolaging.2010.04.024

Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y., 2018. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci. Rep. 8, 6085. https://doi.org/10.1038/s41598-018-24271-9

Chen, M., 2017. MinimalRNN: Toward More Interpretable and Trainable Recurrent Neural Networks. ArXiv171106788 Cs Stat.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J., 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks, in: JMLR Workshop and Conference Proceedings. NIH Public Access, p. 301.

Ding, Y., Sohn, J.H., Kawczynski, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Lituiev, D., Copeland, T.P., Aboian, M.S., Mari Aparici, C., Behr, S.C., Flavell, R.R., Huang, S.-Y., Zalocusky, K.A., Nardo, L., Seo, Y., Hawkins, R.A., Hernandez Pampaloni, M., Hadley, D., Franc, B.L., 2018. A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. Radiology 290, 456–464. https://doi.org/10.1148/radiol.2018180958

Duchesne, S., Caroli, A., Geroldi, C., Collins, D.L., Frisoni, G.B., 2009. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage 47, 1363–1370. https://doi.org/10.1016/j.neuroimage.2009.04.023

Eriksson, D., Bindel, D., Shoemaker, C., 2015. Surrogate optimization toolbox (pysot).

Esteban, C., Staeck, O., Baier, S., Yang, Y., Tresp, V., 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks, in: 2016 IEEE International Conference on Healthcare Informatics (ICHI). Ieee, pp. 93–101.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J.A., Cavallin, L., Guo, C.-J., Ekman, U., Muehlboeck, J.-S., Simmons, A., Barroso, J., Wahlund, L.-O., Westman, E., 2017.

Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. Sci. Rep. 7, 46263. https://doi.org/10.1038/srep46263

García-Laencina, P.J., Sancho-Gómez, J.-L., Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. Neural Comput. Appl. 19, 263–282.

Ghazi, M., Nielsen, M., Pai, A., Cardoso, M.J., Modat, M., Ourselin, S., Sørensen, L., 2019. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. Med. Image Anal. 53, 39–46. https://doi.org/10.1016/j.media.2019.01.004

Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Mach Learn 45, 171–186. https://doi.org/10.1023/A:1010920819831

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2019. Deep Neural Networks and Kernel Regression Achieve Comparable Accuracies for Functional Connectivity Prediction of Behavior and Demographics. bioRxiv 473603. https://doi.org/10.1101/473603

Ilievski, I., Akhtar, T., Feng, J., Shoemaker, C.A., 2017. Efficient Hyperparameter Optimization for Deep Learning Algorithms Using Deterministic RBF Surrogates, in: Thirty-First AAAI Conference on Artificial Intelligence. Presented at the Thirty-First AAAI Conference on Artificial Intelligence.

Ito, K., Ahadieh, S., Corrigan, B., French, J., Fullerton, T., Tensfeldt, T., Alzheimer's Disease Working Group, 2010. Disease progression meta-analysis model in Alzheimer's disease. Alzheimers Dement. J. Alzheimers Assoc. 6, 39–53. https://doi.org/10.1016/j.jalz.2009.05.665

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L.G., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., Weiner, M.W., 2008. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. J. Magn. Reson. Imaging JMRI 27, 685–691. https://doi.org/10.1002/jmri.21049

Jack, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., Lesnick, T.G., Pankratz, V.S., Donohue, M.C., Trojanowski, J.Q., 2013. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol. 12, 207–216. https://doi.org/10.1016/S1474-4422(12)70291-0

Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9, 119–128. https://doi.org/10.1016/S1474-4422(09)70299-6

Johnson, J.K., Gross, A.L., Pa, J., McLaren, D.G., Park, L.Q., Manly, J.J., 2012. Longitudinal change in neuropsychological performance using latent growth models: a study of mild cognitive impairment. Brain Imaging Behav. 6, 540–550. https://doi.org/10.1007/s11682-012-9161-8

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ. 38, 2895–2907.

Kingma, D.P., Ba, L.J., 2015. Adam: A Method for Stochastic Optimization.

Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2019. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cereb. Cortex N. Y. N 1991 29, 2533–2551. https://doi.org/10.1093/cercor/bhy123

Lei, B., Yang, P., Wang, T., Chen, S., Ni, D., 2017. Relational-Regularized Discriminative Sparse Learning for Alzheimer's Disease Diagnosis. IEEE Trans. Cybern. 47, 1102–1113. https://doi.org/10.1109/TCYB.2016.2644718

Li, J., Kong, R., Liegeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A.J., Sabuncu, M.R., Ge, T., Yeo, B.T., 2019. Global Signal Regression Strengthens Association between Resting-State Functional Connectivity and Behavior. https://doi.org/10.1101/548644

Lipton, Z.C., Kale, D.C., Elkan, C., Wetzel, R., 2016a. Learning to diagnose with LSTM recurrent neural networks, in: International Conference on Learning Representations (ICLR). Presented at the International Conference on Learning Representations (ICLR).

Lipton, Z.C., Kale, D.C., Wetzel, R., 2016b. Modeling missing data in clinical time series with rnns. Mach. Learn. Healthc.

Liu, M., Zhang, J., Adeli, E., Shen, D., 2019. Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer's Disease Diagnosis. IEEE Trans. Biomed. Eng. 66, 1195–1206. https://doi.org/10.1109/TBME.2018.2869989

Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S., Alexander, D.C., Consortium, the E., Initiative, for the A.D.N., 2018. TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease. ArXiv180503909 Q-Bio Stat.

Marquand, A.F., Williams, S.C.R., Doyle, O.M., Rosa, M.J., 2014. Full Bayesian multi-task learning for multi-output brain decoding and accommodating missing data, in: 2014 International Workshop on Pattern Recognition in Neuroimaging. Presented at the 2014 International Workshop on Pattern Recognition in Neuroimaging, pp. 1–4. https://doi.org/10.1109/PRNI.2014.6858533

Mazumder, R., Hastie, T., Tibshirani, R., 2010. Spectral regularization algorithms for learning large incomplete matrices. J. Mach. Learn. Res. 11, 2287–2322.

McArdle, J.J., Small, B.J., Bäckman, L., Fratiglioni, L., 2016. Longitudinal Models of Growth and Survival Applied to the Early Detection of Alzheimer's Disease: J. Geriatr. Psychiatry Neurol. https://doi.org/10.1177/0891988705281879

Mohs, R.C., Knopman, D., Petersen, R.C., Ferris, S.H., Ernesto, C., Grundman, M., Sano, M., Bieliauskas, L., Geldmacher, D., Clark, C., Thal, L.J., 1997. Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's Disease Assessment Scale that broaden its scope. The Alzheimer's Disease Cooperative Study. Alzheimer Dis. Assoc. Disord. 11 Suppl 2, S13-21.

Mondal, D., Percival, D.B., 2010. Wavelet variance analysis for gappy time series. Ann. Inst. Stat. Math. 62, 943–966.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Alzheimer's Disease Neuroimaging Initiative, 2015. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage 104, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

Murray, M.E., Graff-Radford, N.R., Ross, O.A., Petersen, R.C., Duara, R., Dickson, D.W., 2011. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. Lancet Neurol. 10, 785–796. https://doi.org/10.1016/S1474-4422(11)70156-9

Nguyen, M., Sun, N., Alexander, D.C., Feng, J., Yeo, B.T.T., 2018. Modeling Alzheimer's disease progression using deep recurrent neural networks, in: 2018 International

Workshop on Pattern Recognition in Neuroimaging (PRNI). Presented at the 2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI), pp. 1–4. https://doi.org/10.1109/PRNI.2018.8423955

Nie, L., Zhang, L., Meng, L., Song, X., Chang, X., Li, X., 2017. Modeling disease progression via multisource multitask learners: A case study with Alzheimer's disease. IEEE Trans. Neural Netw. Learn. Syst. 28, 1508–1519.

Noh, Y., Jeon, S., Lee, J.M., Seo, S.W., Kim, G.H., Cho, H., Ye, B.S., Yoon, C.W., Kim, H.J., Chin, J., Park, K.H., Heilman, K.M., Na, D.L., 2014. Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs. Neurology 83, 1936–1944. https://doi.org/10.1212/WNL.0000000000001003

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., others, 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pham, T., Tran, T., Phung, D., Venkatesh, S., 2017. Predicting healthcare trajectories from medical records: A deep learning approach. J. Biomed. Inform. 69, 218–229. https://doi.org/10.1016/j.jbi.2017.04.001

Popescu, S., Whittington, A., Gunn, R.N., Matthews, P.M., Glocker, B., Sharp, D.J., Cole, J.H., 2019. Nonlinear biomarker interactions in conversion from Mild Cognitive Impairment to Alzheimer's disease. medRxiv 19002378. https://doi.org/10.1101/19002378

Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G.E., Irvine, J., Le, Q., Litsch, K., Mossin, A., Tansuwan, J., Wang, D., Wexler, J., Wilson, J., Ludwig, D., Volchenboum, S.L., Chou, K., Pearson, M., Madabushi, S., Shah, N.H., Butte, A.J., Howell, M.D., Cui, C., Corrado, G.S., Dean, J., 2018. Scalable and accurate deep learning with electronic health records. Npj Digit. Med. 1, 18. https://doi.org/10.1038/s41746-018-0029-1

Regis, R.G., Shoemaker, C.A., 2013. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. Eng. Optim. 45, 529–555.

Rehfeld, K., Marwan, N., Heitzig, J., Kurths, J., 2011. Comparison of correlation analysis techniques for irregularly sampled time series. Nonlinear Process. Geophys. 18, 389–404.

Risacher, S.L., Anderson, W.H., Charil, A., Castelluccio, P.F., Shcherbinin, S., Saykin, A.J., Schwarz, A.J., Alzheimer's Disease Neuroimaging Initiative, 2017. Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline. Neurology 89, 2176–2186. https://doi.org/10.1212/WNL.0000000000004670

Sabuncu, M.R., Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., 2014. Event Time Analysis of Longitudinal Neuroimage Data. NeuroImage 97, 9–18. https://doi.org/10.1016/j.neuroimage.2014.04.015

Sabuncu, M.R., Desikan, R.S., Sepulcre, J., Yeo, B.T.T., Liu, H., Schmansky, N.J., Reuter, M., Weiner, M.W., Buckner, R.L., Sperling, R.A., Fischl, B., 2011. The Dynamics of Cortical and Hippocampal Atrophy in Alzheimer Disease. Arch. Neurol. 68, 1040–1048. https://doi.org/10.1001/archneurol.2011.167

Samtani, M.N., Farnum, M., Lobanov, V., Yang, E., Raghavan, N., DiBernardo, A., Narayan, V., 2012. An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. J. Clin. Pharmacol. 52, 629–644.

Scheltens, P., Blennow, K., Breteler, M.M.B., Strooper, B. de, Frisoni, G.B., Salloway, S., Flier, W.M.V. der, 2016. Alzheimer's disease. The Lancet 388, 505–517. https://doi.org/10.1016/S0140-6736(15)01124-1

Stekhoven, D.J., Bühlmann, P., 2011. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118.

Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S.J., Alzheimer Disease Neuroimaging Initiative, 2010. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage 51, 1405–1413. https://doi.org/10.1016/j.neuroimage.2010.03.051

Sukkar, R., Katz, E., Zhang, Y., Raunig, D., Wyman, B.T., 2012. Disease progression modeling using Hidden Markov Models. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2012, 2845–2848. https://doi.org/10.1109/EMBC.2012.6346556

Sun, N., Mormino, E.C., Chen, J., Sabuncu, M.R., Yeo, B.T.T., 2019. Multi-modal latent factor exploration of atrophy, cognitive and tau heterogeneity in Alzheimer's disease. NeuroImage 201, 116043. https://doi.org/10.1016/j.neuroimage.2019.116043

Suo, Q., Ma, F., Canino, G., Gao, J., Zhang, A., Veltri, P., Agostino, G., 2018. A Multi-Task Framework for Monitoring Health Conditions via Attention-based Recurrent Neural Networks. AMIA. Annu. Symp. Proc. 2017, 1665–1674.

Thung, K.-H., Wee, C.-Y., Yap, P.-T., Shen, D., 2016. Identification of Progressive Mild Cognitive Impairment Patients Using Incomplete Longitudinal MRI Scans. Brain Struct. Funct. 221, 3979–3995. https://doi.org/10.1007/s00429-015-1140-6

Varoquaux, G., 2018. Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage, New advances in encoding and decoding of brain signals 180, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack, C.R., Alzheimer's Disease Neuroimaging Initiative, 2009. MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. Neurology 73, 294–301. https://doi.org/10.1212/WNL.0b013e3181af79fb

Wang, H., Nie, F., Huang, H., Yan, J., Kim, S., Risacher, S., Saykin, A., Shen, L., 2012. High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimer\textquotesingle s Disease Progression Prediction, in: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc., pp. 1277–1285.

Wang, X., Shen, D., Huang, H., 2016. Prediction of Memory Impairment with MRI Data: A Longitudinal Study of Alzheimer's Disease. Med. Image Comput. Comput.-Assist. Interv. MICCAI Int. Conf. Med. Image Comput. Comput.-Assist. Interv. 9900, 273–281. https://doi.org/10.1007/978-3-319-46720-7_32

Wang, X., Sontag, D., Wang, F., 2014. Unsupervised Learning of Disease Progression Models, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14. ACM, New York, NY, USA, pp. 85–94. https://doi.org/10.1145/2623330.2623754

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30, 377–399.

Xie, Q., Wang, S., Zhu, J., Zhang, X., 2016. Modeling and predicting AD progression by regression analysis of sequential clinical data. Neurocomputing, Learning for Medical Imaging 195, 50–55. https://doi.org/10.1016/j.neucom.2015.07.145

Young, A.L., Marinescu, R.V., Oxtoby, N.P., Bocchetta, M., Yong, K., Firth, N.C., Cash, D.M., Thomas, D.L., Dick, K.M., Cardoso, J., Swieten, J. van, Borroni, B., Galimberti, D., Masellis, M., Tartaglia, M.C., Rowe, J.B., Graff, C., Tagliavini, F., Frisoni, G.B., Laforce, R., Finger, E., Mendonça, A. de, Sorbi, S., Warren, J.D., Crutch, S., Fox, N.C., Ourselin, S., Schott, J.M., Rohrer, J.D., Alexander, D.C., 2018.

Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. Nat. Commun. 9, 1–16. https://doi.org/10.1038/s41467-018-05892-0

Zhang, D., Shen, D., 2012. Predicting Future Clinical Changes of MCI Patients Using Longitudinal and Multimodal Biomarkers. PLOS ONE 7, e33182. https://doi.org/10.1371/journal.pone.0033182

Zhang, X., Mormino, E.C., Sun, N., Sperling, R.A., Sabuncu, M.R., Yeo, B.T.T., Initiative, the A.D.N., 2016. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. Proc. Natl. Acad. Sci. 113, E6535–E6544. https://doi.org/10.1073/pnas.1611073113

Zhou, J., Liu, J., Narayan, V.A., Ye, J., 2013. Modeling disease progression via multi-task learning. NeuroImage 78, 233–248. https://doi.org/10.1016/j.neuroimage.2013.03.073

Zhu, Y., Sabuncu, M.R., 2018. A Probabilistic Disease Progression Model for Predicting Future Clinical Outcome. ArXiv180305011 Cs Stat.

Supplementary

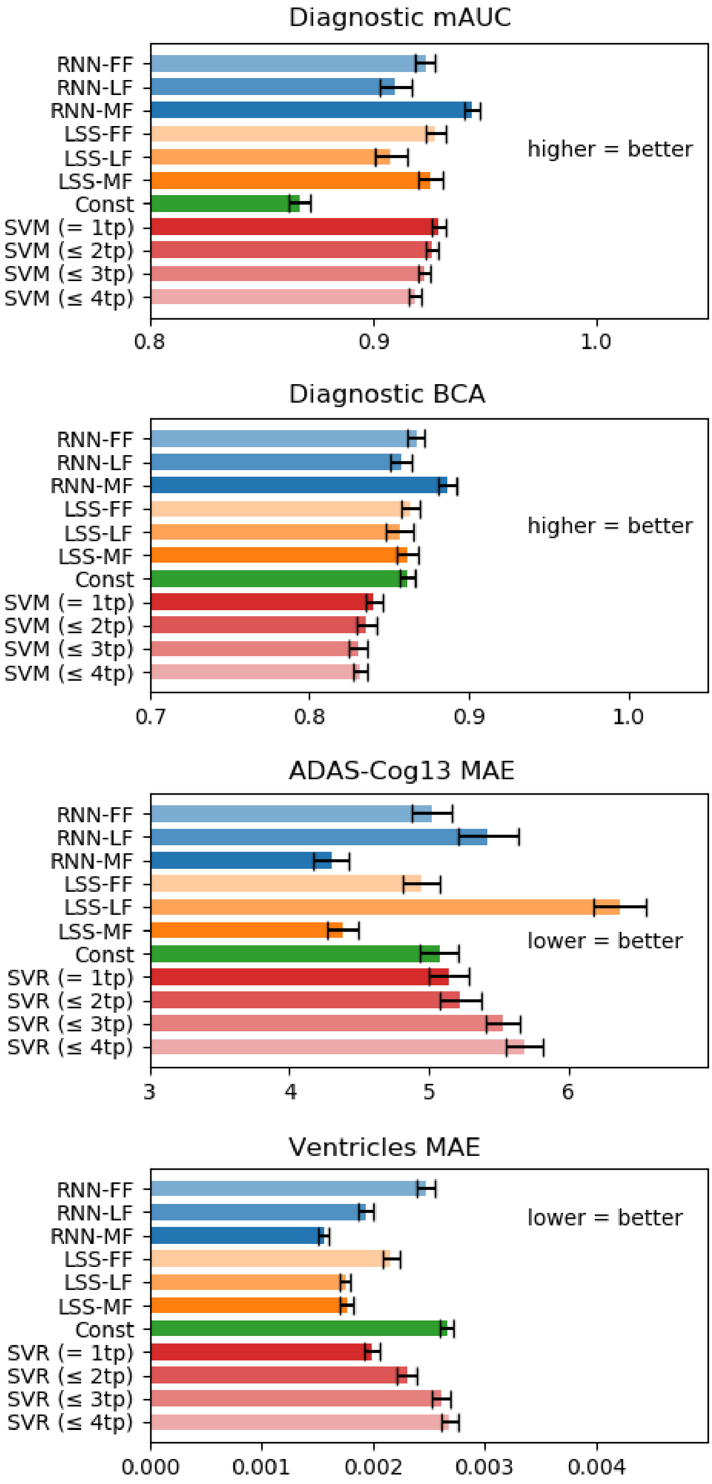


Figure S1. Performance of all models averaged across 20 test sets. Error bars show standard error across test sets. For clinical diagnosis, higher mAUC and BCA values indicate better performance. For ADAS-Cog13 and Ventricles, lower MAE indicates better performance. FF indicates forward filling. LF indicates linear filling. MF indicates model filling. SVM/SVR (= 1tp) utilized one input timepoint. SVM/SVR (≤ 2tp) utilized at most 2 input timepoints (see Section 2.5.2 for details) and so on.

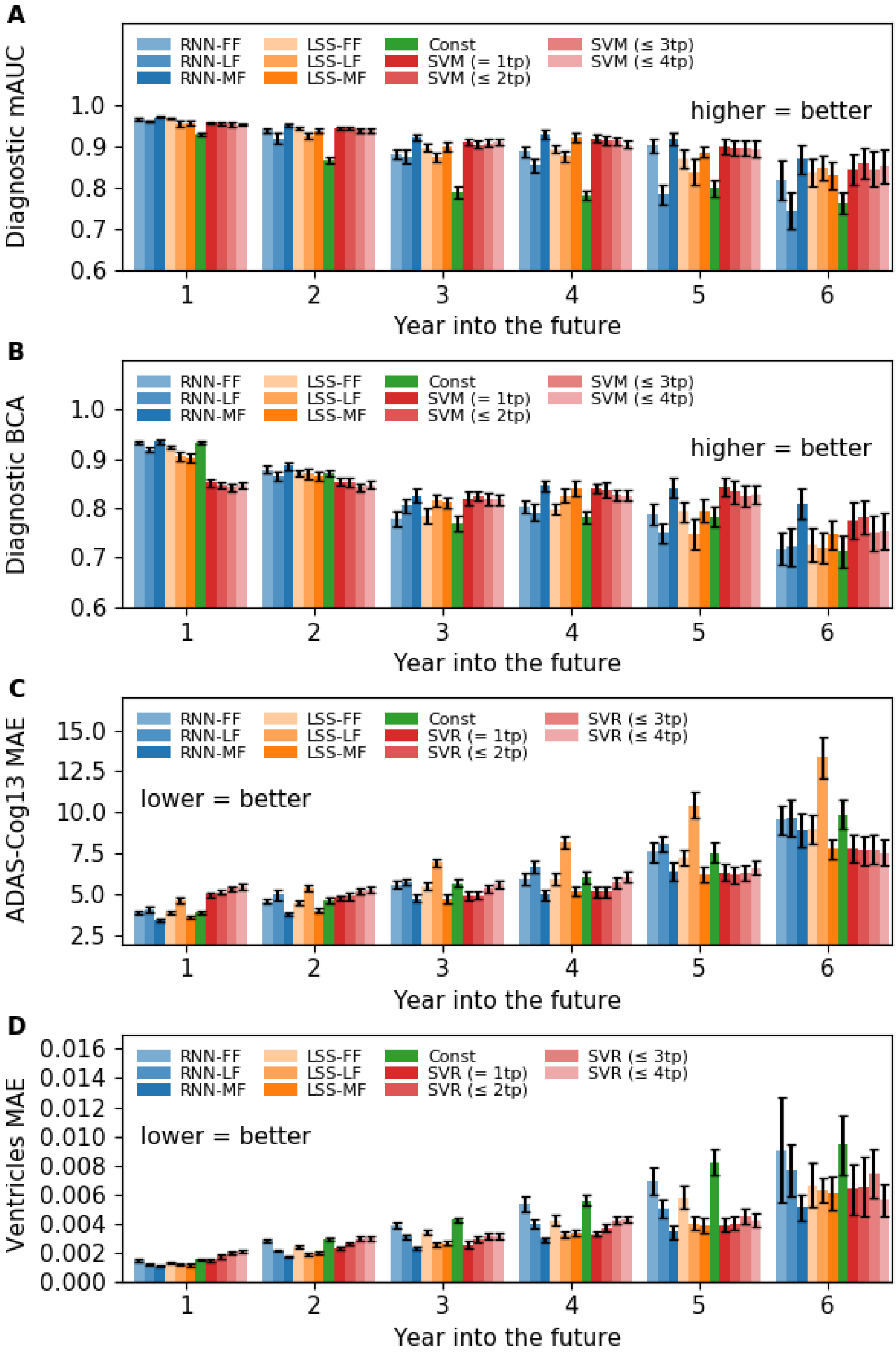


Figure S2. Prediction performance from Figure S1 broken down in yearly interval up to 6 years into the future. For clinical diagnosis, higher mAUC and BCA values indicate better performance. For ADAS-Cog13 and Ventricles, lower MAE indicates better performance. FF indicates forward filling. LF indicates linear filling. MF indicates model filling. SVM/SVR (= 1tp) utilized one input timepoint. SVM/SVR (≤ 2tp) utilized at most 2 input timepoints (see Section 2.5.2 for details) and so on.