Early Detection of Alzheimer's Disease Using Neuropsychological Tests: a Predict-Diagnose Approach Using Neural Networks

Devarshi Mukherji
University of Michigan
Manibrata Mukherji
United Wholesale Mortgage
Nivedita Mukherji (mukherji@oakland.edu)
Oakland University

Research Article

Keywords: Alzheimer's Disease (AD), Neuropsychological Tests, Neural Networks, Predict-Diagnose Approach, neurodegenerative disease

Posted Date: November 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1099058/v1

Early Detection of Alzheimer's Disease using Neuropsychological Tests: A Predict-Diagnose Approach using Neural Networks

by

Devarshi Mukherji

devarshi@umich.edu

University of Michigan, MI

Manibrata Mukherji

manimukh@gmail.com

United Wholesale Mortgage, Pontiac, MI

Nivedita Mukherji

mukherji@oakland.edu

Oakland University, Rochester, MI

Alzheimer's Disease Neuroimaging Initiative[1]

[1]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

Alzheimer's Disease (AD) is the most expensive and currently incurable disease that affects a large number of the elderly globally. One in five Medicare dollars is spent on AD-related tests and treatments. Accurate AD diagnosis is critical but often involves invasive and expensive tests that include brain scans and spinal taps. Recommending these tests for only patients who are likely to develop the disease will save families of cognitively normal individuals and hospitals from unnecessary expenditures. Moreover, many of the subjects chosen for clinical trials for AD therapies never develop any cognitive impairment and prove not to be ideal candidates for those trials. It is thereby critical to find inexpensive ways to first identify individuals who are likely to develop cognitive impairment and focus attention on them for in-depth testing, diagnosing, and clinical trial participation. Research shows that AD is a slowly progressing disease. This slow progression allows for early detection and treatment, but more importantly, gives the opportunity to predict the likelihood of disease development from early indications of memory lapses. Neuropsychological tests have been shown to be effective in identifying cognitive impairment. Relying exclusively on a set of longitudinal neuropsychological test data available from the ADNI database, this paper has developed Recurrent Neural Network (RNN) models to predict future neuropsychological test results and Multi-Level Perceptron (MLP) models to diagnose the future cognitive states of individuals based on those predicted results. The RNNs use sequence prediction techniques to predict test scores for two to four years in the future. The predicted scores and predictions of cognitive states based on them showed a high level of accuracy for a group of test subjects, when compared with their known future cognitive assessments conducted by ADNI. This shows that a battery of neuropsychological tests can be used to track the cognitive states of people above a certain age and identify those who are likely to develop cognitive impairment in the future. This ability to triage individuals into those who are likely to remain normal and those who will develop cognitive impairment in the future, advances the quest to find appropriate candidates for invasive tests like spinal taps for disease identification, and the ability to identify suitable candidates for clinical trials.

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease that affects around 50 million people globally [16]. Projections show that 1 in 85 people will develop AD by 2050. In spite of the scale of the problem, there is no cure for the disease and countless clinical trials have been unsuccessful in finding effective treatment [5, 8]. In 2018, the Alzheimer's Association estimated that the average cost for caring for an Alzheimer's Disease patient is $350,174 – making it the most expensive disease in the United States.

Research shows that AD is an incredibly slowly progressing disease [3, 22, 9] and takes many years from initial cognitive decline to full-blown disease development. Due to this slow progression, early detection of AD can be crucial in both its treatment and prevention. Currently, the detection of AD relies on invasive and expensive tests - spinal taps for CSF Tau protein, brain scans, and blood biomarker detections [19, 12, 18]. These tests contribute to the overall costs associated with AD assessment and treatment mentioned above. It is therefore critical to develop an effective, less expensive, and unintrusive screening method to identify people who are at risk of developing AD so that the expensive and intrusive tests can be used only for those patients. Moreover, multiple ongoing global efforts aim to identify the optimal candidates for their clinical trials (cohorts) to test new therapies and understand disease progression. The cohorts selected for these trials often include patients who do not develop cognitive impairment during their participation and result in an unfortunate waste of resources [18, 7].

It has been shown that neuropsychological tests are effective in the diagnosis of AD and in the identification of patients that are likely to experience AD progression [21, 2, 17, 1, 4, 24, 11]. These tests are much cheaper to administer than CSF spinal taps and brain scans and can be used to identify patients at risk of developing cognitive decline and those that are not. The goal of this study is to use multiple Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs) on a set of well-established neuropsychological tests such as the ones used by [23] to

assess the current cognitive state of a subject and use current and past test score data to predict if a subject is likely to develop cognitive impairment within the next two to four years. This segmentation of subjects into those who are expected to remain cognitively normal and those likely to develop problems provides doctors and researchers an inexpensive and non-intrusive method to identify people that require more invasive interventions and also select appropriate subjects for clinical trials. Access to inexpensive assessments of cognitive state can be of great benefit to people around the world, especially those in underdeveloped countries where low-cost diagnosis methods can be used to preserve critical resources for only the most high-risk patients.

All data for this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. ADNI was launched in 2003 in cooperation with the National Institutes of Health (NIH) and the National Institute on Aging to better understand AD development and discover new therapeutic techniques. This database includes data on neuropsychological tests, along with many other metrics used to monitor AD progression. Based on the research on neuropsychological tests in AD detection and prediction, five different neuropsychological tests (MMSE, ADAS Q4, ADAS Cog11, ADAS Cog13, and FAQ) were chosen for this study - [2] offers a brief description of these tests. The ADNI database includes data on all of these tests.

As mentioned before, this study uses Recurrent Neural Networks (RNNs) for prediction and Multi-Layer Perceptrons (MLPs) for diagnosis. MLPs are a class of Artificial Neural Networks (ANNs) that use multiple layers of neurons and can be used to classify input data. RNNs are machine learning models that use sequences of data to predict future values [14, 6, 10, 13]. RNNs have been used effectively in many domains ranging from language recognition to marketing to healthcare. In the context of AD, [23] is one of the first studies that used RNN in AD prediction. Using NACC data, the paper used 78 features and a global CDR (Clinical Dementia Rating) score to identify patients that are likely to experience AD progression based on changes in CDR scores. Unlike ADNI, the NACC data are not collected at equal intervals for all subjects

necessitating [23] to introduce time between visits as a separate feature. The paper utilized a large number of features (78) in a multi-feature model to identify patients that are likely to experience worsening of the disease in the future. In contrast, the RNNs used in the current study are developed individually for each of the five chosen tests and their output is analyzed using MLPs to diagnose a patient as likely to remain cognitively normal or not in the next two to four years. Consequently, this paper can be viewed as a less expensive step that should be conducted before a full-blown RNN that uses a comprehensive and expensive set of features is used to determine how rapidly a patient's cognitive decline is expected to progress in the future. The objectives of this study are:

1. To develop RNNs, one for each neuropsychological test, that can predict scores using a sequence of four test scores from the past.

2. To develop MLPs that use all the five test scores predicted by the RNNs at a particular visit by a patient after the baseline visit to diagnose whether the patient has cognitive impairment at that visit.

3. To combine the above two sets of neural network models to predict test scores for each test over the next 2-4 years and diagnose upon those predictions.

All of these objectives were met in this study.

2 Methods

2.1 Data

2.1.1 Data, Software, and Packages

All data used in this project were obtained from the ADNI database. ADNI is a longitudinal study launched in 2003 in cooperation with the National Institutes of Health (NIH) and the National Institute on Aging. ADNI uses adult volunteers between the ages of 55 and 90. The initial cohort included 200 cognitively normal (CN) subjects, 400 subjects with mild cognitive impairment (MCI), and 200 subjects diagnosed with Alzheimer's Disease (AD). Each individual is assigned a unique identification number or RID. CN subjects have a Clinical Dementia Rating (CDR) of 0 and MMSE score between 24 and 30; MCI subjects show signs of memory loss, have a CDR score of 0.5, and are at least one standard deviation below the mean score on the delayed recall portion of the Wechsler Memory Scale's Logical Memory II. Those who are diagnosed with AD are diagnosed in accordance with the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [15]. The ADNI study has completed three phases to date and is currently in its ADNI3 phase. The first phase, ADNI1, started in October 2004 and spanned five years. ADNI-GO was the second phase and lasted two years after its launch in September 2009, and ADNI2 spanned five years since its launch in September 2011. In each phase, new participants were recruited, and participants from the previous phases continued. The ADNI dataset contains demographic data of the subjects as well as results of various neuropsychological tests, results of brain scans, and other physiological tests conducted at various intervals. The relevant data for this study are obtained from a subset of the ADNI dataset called ADNIMERGE and was downloaded on May 14, 2019. Data extraction and processing were conducted using the R programming language running in RStudio version 1.2. The recurrent neural network analysis was conducted using the Keras package version 2.2.5.0 running in RStudio. Descriptive statistical

analysis was conducted using the statistical software package Stata 14.

Since the objectives of this paper are based on RNNs that use past neuropsychological test scores to predict future test scores, it is important to select subjects who have sufficient time series data for the chosen tests. ADNI subjects undergo an initial assessment that determines their baseline cognitive states and then have follow-up tests at months 6, 12, 18, 24, 36, 48, 60, 72, 84, and so on. The data for this study are primarily derived from the baseline, months 6, 12, 24, 36, 48, 60, and 72. (Months missing in this interval, such as 18, do not have any neuropsychological test values in the ADNI database.) A sequence of length four uses the baseline, and months 6, 12, and 24 data; a sequence of length six also includes months 36 and 48, and a sequence of length eight includes months 60 and 72.

2.1.2 Neuropsychological Test Selections

Five neuropsychological test scores were used to assess the cognitive state of the subjects. The tests are MMSE, ADAS Q4, ADAS Cog-11, ADAS Cog-13, and FAQ. These tests have been used to demonstrate their effectiveness in diagnosis and prediction of cognitive impairment [2]. Additionally, neuropsychological tests were used in a regression model to determine the impact on change in CDR sum of scores [25] and "MMSE, FAQ, and ADAS-cog were identified as prognostic factors to detect cognitive decline in CDR-SB." Other test scores, such as the Wechsler Logical Memory Delay (LDELTOTAL), are also available. However, the number of missing values of the latter test was greater than for the other tests, and the test was given less frequently (every two years as opposed to half-yearly or yearly) than the other tests, therefore limiting the number of subjects that could be used for the final test of the models, taking it out of consideration for this study.

2.2 Training, Validation and Test Data Set Creation

As noted earlier, the battery of neuropsychological tests is administered during scheduled visits that occur at the initial visit (baseline) and then at months 6, 12, 24, and beyond. All participants do not attend all of the scheduled visits, and even when they do, results for all scheduled tests are not available. Consequently, there are many missing values for all of the tests that are part of the ADNI study. Since the goal of this research is to start with at least four data points for a subject and predict the status in the next two to four years, sequences of length 4, 6, and 8 were extracted for each of the five neuropsychological tests separately from the ADNI1, ADNI-GO, and ADNI2 datasets and combined to form the training and validation data for the test-score prediction RNN models. No special care was taken to collect sequences for the same patient across the neuropsychological tests since the generic pattern of disease progression is what the RNNs need to learn, not the progression pattern of the same patient across the tests.

Special care had to be taken to re-create or impute missing values. To use mostly real data values and reduce the number of imputations, it was ensured that subjects had no more than 2 missing test scores. Moreover, subjects who had a missing value on the last visit of a sequence of length 4, 6, or 8 were also dropped. The latter constraint ensured that no model would learn the generic trend of disease progression using imputed values as the most recent data point. A simple average method was used to impute missing values. As an example, for a model of length four, a subject with missing data for month 12 was assigned for month 12, the average of the baseline, month 6, and month 24 values. It is one of the many ways in which missing data can be imputed. This imputation method is advantageous since this method preserves the average of the sequence.

Imputed sequences of length 6 from the combined dataset were used to generate the training and validation data for the 5th and 6th value prediction RNN models. For example, for the 5th value prediction RNN model, the model was trained with sequence values 1 through 4 as

input data and sequence value 5 as the output data value. The model learned how the 5th value was associated with the previous four values using the training data and tested what it learned on the validation data to determine its accuracy. Similarly, for the 6th value prediction model, sequence values 1 through 5 were used as the input data and value 6 was used as the output data value. Eighty percent of the rows were used to create the training set and the remaining 20% of rows made up the validation set. For the 7th and 8th element prediction models, a similar set of steps were followed using the combined data set containing sequences of length 8.

The MLPs need to learn to diagnose the cognitive state of a single patient given the predicted test score values of the five neuropsychological tests of that patient. Sequences of length 4, 6, and 8 were extracted for a given RID for each of the five neuropsychological tests separately from the ADNI1, ADNI-GO, and ADNI2 datasets, yielding 15 different collections. The ADNI diagnosis values (CN, MCI, or AD) were also saved for each RID and for each sequence at the 4th, 6th, and 8th visits from the three datasets (ADNI1, ADNI-GO, and ADNI2). If the test score of any of the five tests at visit 4, 6 or 8 and the corresponding ADNI diagnosis was not available, the RID was removed from the dataset. Among the RIDs that remained, if any sequence for the five tests contained more than two unavailable scores, the RID was removed from the dataset. The remaining test scores formed the core for generating the training and validation data for the diagnosis MLP models. A classification output value of 0, cognitively normal (CN), was generated for an ADNI diagnosis value of CN and a classification output value of 1, non-CN, was generated for an ADNI diagnosis values of MCI and AD. For each sequence length, the data was then sorted according to the classification value (0 or 1), and 80% of the smaller of the two classified sets was combined with a matching number of sequences from the other set of sequences to create the training set. This procedure was carefully incorporated due to the models' enhanced ability to learn if there are an equal number of 0s and 1s. Finally, the remaining sequences were pulled together to form the validation data set.

To determine if a combination of neuropsychological tests can be used to predict the future cognitive state of an individual, it is necessary to create a test dataset of individuals for whom longitudinal data of length 8 (baseline to 72 months) are available for all of the five tests chosen for this study. Starting with sequences of length four, these individuals' test scores would be predicted for the next two and four years using the RNN models of sequence prediction and those predictions would be collectively used to determine their future cognitive state using the diagnosing MLPs. Then, the predicted outcomes would be compared with the actual ADNI diagnosis to determine the accuracy of the predictions. After applying the same criterion of no more than two missing values, no missing diagnosis value, and ensuring that these sequences were not used as training or validation data for any RNN or MLP model, only 66 patients who had data for all the five neurological tests could be obtained from the oldest cohort – ADNI1. Descriptive statistics for this cohort is given in Table 1.

Table 2 provides the range of test score values for each test, along with their cutoff scores for normal values which were obtained from the different references discussed in sections 1 and 2. The last column of the table provides the normalization calculations used for each test to convert the data range from 0 to 1. Even though the cutoff values were not used to determine the classification output values of the MLP models, these cutoff values were used to determine the efficacy of the sequence prediction RNN models and to determine the contribution of the normal and abnormal patients towards the overall diagnosis as discussed in a later section.

2.2.1 Model Creation and Application

Each diagnosis and sequence prediction model was trained using its training data and validated using its validation data. The loss curve was plotted for each model, and both the model and the loss curve were saved. The models with the highest prediction accuracy were saved for later use in the test data prediction phase.

Table 3 gives the number of unique RIDs used to train and validate the sequence prediction

models. Table 4 gives the number of unique RIDs used to train and validate the 4th step, 6th step, and 8th step diagnosis models. The numbers show that for each test, the number of observations for the datasets decreases when going from sequences of length 6 to sequences of length 8 since the number of RIDs that have missing values over a 72-month period is much higher than it is over a 48-month period.

Figure 1 shows the structure of three neural network models that were used in this study. Figure 1a shows the structure of the diagnosis MLP that was used to diagnose the five neurological test scores at month 48. The neural network has an input layer, a "sigmoid" output later, and four hidden layers. Figure 1b shows the structure of the diagnosis MLP that was used to diagnose the five neurological test scores at month 72. It also has an input layer, a "sigmoid" output later, and four hidden layers but uses different numbers of neurons at each hidden layer. The 'relu' activation was used for every non-output layer and the "$mean_squared_error$" loss function and the 'adam' optimizer was used to compile the models. Figure 1c shows the common structure of the sequence prediction RNNs that were used to predict the five neurological test scores. Two layers of LSTM RNNs were used followed by a dense output layer. The 'relu' activation was used for every layer and the "$mean_squared_error$" loss function and the 'adam' optimizer was used to compile the models. The main goal of this study lies in the combination of the diagnostic and predictive models. The combination of the models is used on the final test data created for each neuropsychological test. The final test data set is the set of RIDs that is used to test the accuracy and practicality of the study, and it consists of 66 patients who all have test scores for each of the neuropsychological tests. The first four values of each of the sequences were fed into the best 5th-element and 6th-element sequence prediction models, which predicted the 5th and 6th values of the sequence (two years into the future in relation to value 4), respectively. The best diagnosis prediction model was then used to diagnose the CN-non CN status of each RID using the predicted 6th values of all the five tests for that RID (diagnosis two

years after the 4th value). The best 7th-element and 8th-element sequence prediction models were then used in order to predict the 7th and 8th values of each of the sequences, respectively, starting with the sequences of length six that was used in the previous step. The best diagnosis prediction model was then used to diagnose the CN-non CN status of each RID using the predicted 8th values of all the five tests for that RID (diagnosis four years after the 4th value). Then these diagnosis predictions, two and four years after the fourth value, were compared to the real diagnosis of the RIDs given by ADNI (using DX values) at those respective times to determine the accuracy of the study as a whole.

3 Results

There are three sets of results for this work, and they relate to 1) accuracies of the diagnosis of cognitive states 2) how closely the predicted values of the test scores follow the trend of the actual values in the validation data set two years (6th value) and four years (8th value) after month 24, and 3) accuracies of the diagnosis when the diagnosis MLPs and the sequence prediction RNNs are combined to predict on the test dataset.

3.1 Diagnosis Models' Accuracies

The diagnosis models learn from the training data to determine which combination of the five test scores should be assigned a CN (0) value and which sequences should be assigned a non-CN (1) value. When the trained models are applied to the validation data, the models' assignments of 0 and 1 are compared with the labeling of 0 (ADNI diagnosis CN, DX = 1) and 1 (ADNI diagnosis MCI/AD, DX = 2/3) of the validation data sequences to arrive at the accuracy results.

Table 5 gives the accuracies of the diagnosis models. The diagnosis at month 48 is 78.85% accurate and the diagnosis at month 72 is 77.88% accurate. The accuracy values were determined by taking the proportion of correct diagnoses to the total number of diagnoses.

The boxplots of Figure 2 and Figure 4 serve as a corroboration of the accuracies noted above.

The plots also corroborate the accuracy of the normal-abnormal cutoff values shown in Table 2. The boxplots for each graph are split into two groups, CN and Non-CN. CN depicts the range of scores diagnosed as 0 by the diagnosis models, and Non-CN depicts those diagnosed as 1. The stark contrast in the ranges of the CN and Non-CN boxplots shows that the diagnosis models were able to effectively learn for all of the five tests.

An aspect of the boxplots of Figure 2 that deserves special mention is that even though the prediction accuracies are in the high seventies, there are still some RIDs that are incorrectly classified. For example, for MMSE (Figure 2a), The CN group contains some RIDs that are in the abnormal range (< 28) and the Non-CN group contains some RIDs that are in the normal range ($>= 28$). Figure 3a shows a bar graph for all the five tests in which the CN and Non-CN groups are further split into normal and abnormal percentage values based on the cutoff values for the test from Table 2. For example, for MMSE, the CN group contains 88% RIDs that are in the normal range and 12% RIDs that are in the abnormal range. It is these 12% abnormal RIDs which run the risk of being mis-diagnosed as normal and hence dropped from further monitoring and treatment. Similarly, for MMSE, the Non-CN group contains 31% RIDs that are in the normal range and 69% RIDs that are in the abnormal range. Once again, it is the 31% normal RIDs that run the risk of being monitored and treated causing a waste of valuable resources. Such mis-diagnosis is inevitable in the context of machine learning approaches, especially when training data is scarce. But one important observation that should be made in regards to the diagnosis reached at by ADNI, as shown in Figure 3b, is that the mix of normal-abnormal mis-diagnosis percentages are not any better when we consider the validation data and group them by the ADNI DX value (CN and Non-CN, that is, MCI/AD). As seen in Figure 3b, for MMSE, say, the Non-CN group contains 43% RIDs that are in the normal range (compared to 27% using the predicted diagnosis of our models) and run the risk of being monitored and treated causing a waste of valuable resources. All the tests in Figure 3a, when compared with the corresponding

tests in Figure 3b shows that our prediction and diagnosis models perform better in reducing the mis-diagnosis percentage values thereby promising to be a cost effective solution for initial screening of Alzheimer's patients.

Figures 5a and 5b similarly capture the mis-diagnosis percentage values for predicted diagnosis and ADNI diagnosis outcomes at month 72 and it is clear once again that our prediction and diagnosis models perform better in reducing the mis-diagnosis percentage values.

3.2 Performance of the Sequence Prediction Models

Figure 6 consists of a series of graphs that demonstrate the performance of the sequence prediction models when predicting the 6th value based on the first four actual values and the predicted fifth value. The validation datasets were chosen for these figures. The blue line depicts the actual 6th value of each of the subjects, while the red line depicts the prediction of the 6th value using the best sequence prediction model. The horizontal yellow line represents the threshold that differentiates test scores defined as CN or non-CN from Table 2. The comparisons show that even if the predicted values are not exactly identical to the actual values, they are close and more importantly, follow the general trend of the actual data. If the predicted values did not follow the general trend, the diagnoses based on these values would not be accurate. Moreover, the inaccuracy is of greater concern if the predicted and actual values lied on two different sides of the cutoff line.

Figure 7 displays the performance of the sequence prediction models when predicting the 8th value, the 72nd month, based on the first four actual values and the predicted fifth, sixth, and seventh values. In this case as well, the data are for the individuals in the validation datasets. The results are similar to those reported above for predicting the 6th value.

Figures 8 and 9 display results, similar to those in Figures 6 and 7, for the 66 individuals in the test dataset. The results depicted in Figure 8 for predictions two years ahead, show that the percentage of predicted scores on the same side of the cutoff lines are high for all tests for

month 48. The same is true for month 72 as shown in Figure 9.

3.3 Combined Model's Accuracy

The key objective of this study is to determine if neuropsychological test scores can be used to predict if an individual will remain cognitively normal in the next two to four years. If the models, when combined together, predict that a patient will not remain CN over the next two to four years, then those individuals should be monitored more frequently and are likely to be recommended to undergo further testing.

Tables 6 and 7 collectively show that the combined model, when applied to the test data set, is successful in predicting that an individual will be cognitively normal at the end of the next two years with 84.62% accuracy. This accuracy drops to 83.33% for predictions that are four years in the future. Similarly, the combined model is successful in predicting that an individual will be cognitively abnormal at the end of the next two years with 78.9% accuracy. This accuracy rises to 82.8% for predictions that are four years in the future. The subjects predicted to be cognitively abnormal are good candidates for more invasive testing and are likely to serve as good candidates for clinical trials for AD treatments. Hence, even though the ADNI diagnosis is not solely based on these five neuropsychological tests, our technique shows that if we use just the outcomes of these five tests, we can determine with a high degree of accuracy how the disease will progress for patients for whom there are no other data available to aid the diagnosis.

The boxplots of Figure 10 and Figure 12 serve as a corroboration of the accuracies noted above. The plots also corroborate the accuracy of the normal-abnormal cutoff values shown in Table 2. The boxplots for each graph are split into two groups, CN and Non-CN. CN depicts the range of scores diagnosed as 0 by the diagnosis models, and Non-CN depicts those diagnosed as 1. The stark contrast in the ranges of the CN and Non-CN boxplots shows that the diagnosis models were able to predict very effectively on data that it had not seen before during training and validation.

Figure 11a shows a bar graph for all the five tests in which the CN and Non-CN groups are further split into normal and abnormal percentage values based on the cutoff values for the test from Table 2. For example, for MMSE, the CN group contains 98% RIDs that are in the normal range and 2% RIDs that are in the abnormal range. Similarly, for MMSE, the Non-CN group contains 41% RIDs that are in the normal range and 59% RIDs that are in the abnormal range. As before, in regards to the diagnosis reached at by ADNI, as shown in Figure 11b, the mix of normal-abnormal mis-diagnosis percentages are not any better when we consider the test data and group them by the ADNI DX value (CN and Non-CN, that is, MCI/AD). As seen in Figure 11b, for MMSE, say, the Non-CN group contains 57% RIDs that are in the normal range (compared to 41% using the predicted diagnosis of our models) and run the risk of being monitored and treated causing a waste of valuable resources. All the tests in Figure 11a, when compared with the corresponding tests in Figure 11b shows that our prediction and diagnosis models perform better in reducing the mis-diagnosis percentage values thereby promising to be a cost effective solution for initial screening of Alzheimer's patients.

Figures 12 and 13 similarly capture the mis-diagnosis percentage values for predicted diagnosis and ADNI diagnosis outcomes at month 72 and it is clear once again that our prediction and diagnosis models perform better in reducing the mis-diagnosis percentage values.

4 Discussion and Conclusion

This study has demonstrated that it is possible to use machine learning tools on past neuropsychological test scores to predict future values of those tests and predict the future cognitive states of individuals. Using data from hundreds of subjects who participated in the ADNI project, this study used RNN techniques to predict future values of their tests and used MLPs to diagnose individuals as CN or non-CN at a future date. For a cohort of 66 test subjects, the test scores of all five tests were combined to generate a prediction of their future cognitive states based

on the combined results. These results were then matched with the actual CN or MCI/AD status assigned by ADNI at future points in time. The results show that individuals who are predicted by the model to continue to remain CN in all of the tests are highly likely to remain cognitively normal in real life over the next two to four years. Individuals who are predicted to have test scores outside of normal ranges are likely to experience cognitive impairment in the future. These individuals should be monitored and treated.

The analysis of this paper focused on five tests for which consistent data over at least a 72 month period was available for a large number of individuals. Studies such as [2] used a battery of tests to determine which ones are most effective in classifying patients accurately with small differences in CDR scores. Their Figure 2 shows LDELTOTAL, LIMMTOTAL, Q4, TOTALMOD, Q1, MMSESCORE, TOTAL11, FAQTOTAL, and CATVEGESC to be some of the effective tests. This study used Q4 (ADAS Q4), TOTALMOD (ADAS 13), MMSE, TOTAL11 (ADAS 11), and FAQ from this list. The rest of the tests did not have sufficient longitudinal data to be used. Another study, [20], assessed different neuropsychological tests to determine which have good prediction power. While many of the tests identified by these studies had some data available in ADNI, the limitation of having no more than 2 missing values restricted the number of tests that could be considered for this study. The combined model's structure allows one to introduce any test for which sufficient longitudinal data is available from ADNI or any other data source such as NACC (National Alzheimer's Coordinating Center).

The benefit of developing a tool based on neuropsychological test scores that can predict with high accuracy the likelihood of an individual experiencing cognitive difficulties in the next few years derives from the low cost of administering these tests. If these tests become routine for individuals above a certain age or who have some risk factors for developing cognitive impairment, longitudinal performance data can be used with a high level of accuracy to determine which patients will require close monitoring. As treatments for Alzheimer's Disease continue to

develop, this ability to determine who will require close monitoring can allow more invasive and expensive tests to be reserved for them. This can also allow for early intervention, which can be crucial in treatment or prevention of the disease.

5 Declarations

5.1 Ethics approval and consent to participate

ADNI provided access to all data used in this study and the authors used their guidelines for acknowledging them in the author section. The manuscript received approval from ADNI for submission of the research for publication. All authors of the study have provided consent for submitting this research for publication.

5.2 Availability of Data

Data used in this project are available at https://www.kaggle.com/mukherji/datasets?campaign=50c1a4cf-95ba-4b4c-9c0c-2386bde2d81e

5.3 Competing Interests

The authors declare that there are no competing interests.

5.4 Funding

The authors received no external funding to conduct this research.

5.5 Authors' Contributions

DM and MM contributed equally to this work. NM contributed to the formal analysis, visualization, and manuscript preparation.

5.6 Acknowledgements

The authors acknowledge ADNI for allowing them access to the data that is used to conduct this research.

5.7 Authors' Information

This research combines the various interests and expertise of the authors. The main research interest that has motivated the work stems from DM's interest in AD as a student of Neuroscience. As such, DM contributed to the main research concept and methodology of this project. MM's expertise as a computer scientist was critical for the machine learning and neural network programming aspects of the project. NM's expertise in econometrics as a professor of economics was helpful for the statistical analysis and graphical representations of the paper. All authors contributed equally in writing the manuscript.

References

[1] **Albert, Marilyn S., Mark B. Moss, Rudolph Tanzi, and Kenneth Jones**, "Preclinical prediction of AD using neuropsychological tests," *Journal of the International Neuropsychological Society*, Jul 2001, *7* (5), 631–639.

[2] **Battista, Petronilla, Christian Salvatore, and Isabella Castiglioni**, "Optimizing Neuropsychological Assessments for Cognitive, Behavioral, and Functional Impairment Classification: A Machine Learning Study," *Behavioural neurology*, 2017, *2017*, 1–19.

[3] **Berti, Valentina, Cristina Polito, Gemma Lombardi, Camilla Ferrari, Sandro Sorbi, and Alberto Pupi**, "Rethinking on the concept of biomarkers in preclinical Alzheimer's disease," *Neurological Sciences*, May 2016, *37* (5), 663–672.

[4] **Bondi, Mark W., Amy J. Jak, Lisa Delano-Wood, Mark W. Jacobson, Dean C. Delis, and David P. Salmon**, "Neuropsychological Contributions to the Early Identification of Alzheimer's Disease," *Neuropsychology review*, 2008, *18* (1), 73–90.

[5] **Brookmeyer, Ron, Elizabeth Johnson, Kathryn Ziegler-Graham, and H. Michael Arrighi**, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's & dementia*, 2007, *3* (3), 186–191.

[6] **Choi, Edward, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun**, "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks," *JMLR workshop and conference proceedings*, Aug 2016, *56*, 301–318.

[7] **Cummings, Jeffrey, Garam Lee, Travis Mortsdorf, Aaron Ritter, and Kate Zhong**, "Alzheimer's disease drug development pipeline: 2017," *Alzheimer's & dementia : translational research & clinical interventions*, 2017, *3* (3), 367–384.

[8] **Cummings, Jeffrey L., Travis Morstorf, and Kate Zhong**, "Alzheimer's disease drug-development pipeline: few candidates, frequent failures," *Alzheimer's research & therapy*, 2014, *6* (4), 37.

[9] **Dubois, Bruno and Harald Hampel et.al.**, "Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria," *Alzheimer's & dementia*, 2016, *12* (3), 292–323.

[10] **Esteban, Cristóbal, Oliver Staeck, Yinchong Yang, and Volker Tresp**, "Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks," Feb 8, 2016.

[11] **Gainotti, Guido, Davide Quaranta, Maria Gabriella Vita, and Camillo Marra**, "Neuropsychological Predictors of Conversion from Mild Cognitive Impairment to Alzheimer's Disease," *Journal of Alzheimer's disease*, 2014, *38* (3), 481–495.

[12] **Goudey, Benjamin, Bowen J. Fung, Christine Schieber, and Noel G. Faux**, "A blood-based signature of cerebrospinal fluid A1–42 status," *Scientific reports*, 2019, *9* (1), 4163.

[13] **Lipton, Zachary C., David C. Kale, Charles Elkan, and Randall Wetzel**, "Learning to Diagnose with LSTM Recurrent Neural Networks," Nov 11, 2015.

[14] **___ , John Berkowitz, and Charles Elkan**, "A Critical Review of Recurrent Neural Networks for Sequence Learning," May 29, 2015.

[15] **McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan**, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease," *Neurology*, Jul 1, 1984, *34* (7), 939.

[16] **Report, 2018**, "2018 Alzheimer's disease facts and figures," *Alzheimer's & dementia*, Mar 2018, *14* (3), 367–429.
ΩÉrica Maria Lima Pimentel

[17] **Érica Maria Lima Pimentel**, "Role of neuropsychological assessment in the differential diagnosis of Alzheimer's disease and vascular dementia," *Dementia & neuropsychologia*, Jul 2009, *3* (3), 214–221.

[18] **Sevigny, Jeff, Joyce Suhy, Ping Chiao, Tianle Chen, Gregory Klein, Derk Purcell, Joonmi Oh, Ajay Verma, Mehul Sampat, and Jerome Barakos**, "Amyloid PET Screening for Enrichment of Early-Stage Alzheimer Disease Clinical Trials: Experience in a Phase 1b Clinical Trial," *Alzheimer disease and associated disorders*, Feb 15, 2016, *30* (1), 1–7.

[19] **Shi, Liu, Alison L. Baird, Sarah Westwood, Abdul Hye, Richard Dobson, Madhav Thambisetty, and Simon Lovestone**, "A Decade of Blood Biomarkers for Alzheimer's Disease Research: An Evolving Field, Improving Study Designs, and the Challenge of Replication," *Journal of Alzheimer's disease*, 2018, *62* (3), 1181–1198.

[20] **Tabert, Matthias H., Jennifer J. Manly, Xinhua Liu, Gregory Pelton, Sara Rosenblum, Marni Jacobs, Diana Zamora, Madeleine Goodkind, Karen L. Bell, Yaakov Stern, and Devangere P. Devanand**, "Neuropsychological Prediction of Conversion to Alzheimer Disease in Patients With Mild Cognitive Impairment," *JAMA psychiatry (Chicago, Ill.)*, Jun 30, 2017.

[21] **Tierney, M. C., C. Yao, A. Kiss, and I. McDowell**, "Neuropsychological tests accurately predict incident Alzheimer disease after 5 and 10 years," *Neurology*, Jun 13, 2005, *64* (11), 1853–1859.

[22] **Villemagne, Victor L., Samantha Burnham, Pierrick Bourgeat, Belinda Brown, Kathryn A. Ellis, Olivier Salvado, Cassandra Szoeke, S. Lance Macaulay, Ralph Martins, Paul Maruff, David Ames, Christopher C. Rowe, and Colin L. Masters**, "Amyloid deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study," *Lancet neurology*, 2013, *12* (4), 357–367.

[23] **Wang, Tingyan, Robin G. Qiu, and Ming Yu**, "Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks," *Scientific reports*, 2018, *8* (1), 9161–12.

[24] **Weakley, Alyssa, Jennifer A. Williams, Maureen Schmitter-Edgecombe, and Diane J. Cook**, "Neuropsychological test selection for cognitive impairment classification: A machine learning approach," *Journal of clinical and experimental neuropsychology*, 2015, *37* (9), 899–916.

[25] **Yagi, Takuya, Michio Kanekiyo, Junichi Ito, Ryoko Ihara, Kazushi Suzuki, Atsushi Iwata, Takeshi Iwatsubo, and Ken Aoshima**, "Identification of prognostic factors to predict cognitive decline of patients with early Alzheimer's disease in the Japanese Alzheimer's Disease Neuroimaging Initiative study," *Alzheimer's & dementia : translational research & clinical interventions*, 2019, *5* (1), 364–373.

Table 1: **Descriptive Statistics of the 66 Test Dataset Subjects**

Age	#	Gender	#	Education	#	Race	#	Marital Status	#
Age	#	Gender	#	Education	#	Race	#	Marital Status	#
58-68	8	Female	29	8-11	3	Asian	2	Divorced	4
68-78	41	Male	37	12-15	22	Black	4	Married	49
78-88	17			16-20	41	White	60	Never married	1
								Widowed	12

Table 2: **Test Score Cutoffs and Normalizations**

Test	Normal Cutoff	Min	Max	Normalization
MMSE	>=28	0	30	Value/30
ADAS4	<=5	0	10	Value/10
ADAS11	<=10	0	70	Value/70
ADAS13	<=13	0	85	Value/85
FAQ	<=2	0	30	Value/30

Table 3: **Number of Observations for Training and Validation of Sequence Prediction RNNs**

Test	Training 36th & 48th Months	Training 60th & 72nd Months	Validation 36th & 48th Months	Validation 60th & 72nd Months
MMSE	458	230	115	57
ADAS4	458	220	115	55
ADAS11	441	202	110	50
ADAS13	437	209	109	52
FAQ	466	250	116	63

Table 4: **Number of Observations for Training and Validation of Diagnosis MLPs**

Diagnosis Time	Training	Validation
Month 24	632	156
Month 48	428	156
Month 72	264	104

Table 5: **Diagnosis Models' Accuracies**

Diagnosis Time	Accuracy
Month 24	85.25%
Month 48	78.85%
Month 72	77.88%

Table 6: **Combined Model's Prediction Accuracy by Category at Month 48**

Combined Model's Predictions		**ADNI's Assessment**			**Accuracy**
Diagnosis	No. of Subjects	CN	Non-CN	NA	Accuracy
CN	44	33	6	5	84.62%
Non-CN	22	4	15	3	78.9%

Table 7: **Combined Model's Prediction Accuracy by Category at Month 72**

Combined Model's Predictions		**ADNI's Assessment**			**Accuracy**
Diagnosis	No. of Subjects	CN	Non-CN	NA	Accuracy
CN	37	30	6	1	83.33%
Non-CN	29	5	24	0	82.8%

Table 8: **Combined Model's Overall Diagnosis Accuracies**

Diagnosis Time	Accuracy
Month 48	82.76%
Month 72	83.08%

Figure 1: Summary of Diagnosis MLPs and Sequence Prediction RNN Models

(a) Diagnosis MLP for Month 48

Model: "sequential_143"

Layer (type)	Output Shape	Param #
dense_456 (Dense)	multiple	30
dense_455 (Dense)	multiple	60
dense_454 (Dense)	multiple	220
dense_453 (Dense)	multiple	210
dense_452 (Dense)	multiple	55
dense_451 (Dense)	multiple	6

Total params: 581
Trainable params: 581
Non-trainable params: 0

(b) Diagnosis MLP for Month 72

Model: "sequential_410"

Layer (type)	Output Shape	Param #
dense_2177 (Dense)	multiple	30
dense_2176 (Dense)	multiple	300
dense_2175 (Dense)	multiple	1275
dense_2174 (Dense)	multiple	130
dense_2173 (Dense)	multiple	30
dense_2172 (Dense)	multiple	6

Total params: 1,771
Trainable params: 1,771
Non-trainable params: 0

(c) Sequence Prediction RNN

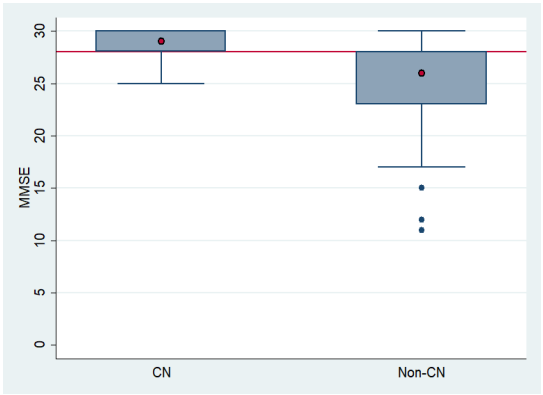
Model: "sequential_17"

Layer (type)	Output Shape	Param #
lstm_34 (LSTM)	(None, 5, 100)	40800
lstm_35 (LSTM)	(None, 25)	12600
dense_17 (Dense)	(None, 1)	26

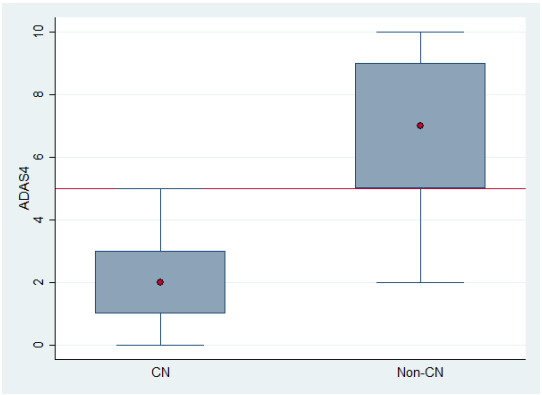
Total params: 53,426
Trainable params: 53,426
Non-trainable params: 0

Figure 2: Ranges of Validation Data Split by the Predicted Diagnosis (CN or Non-CN) Assigned by the Diagnosis Model at Month 48 (with Normal Cutoffs)

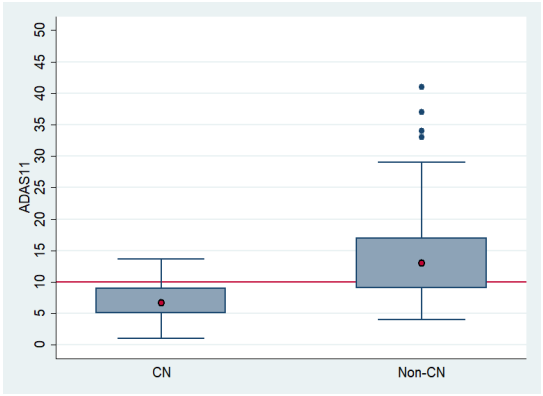
(a) MMSE



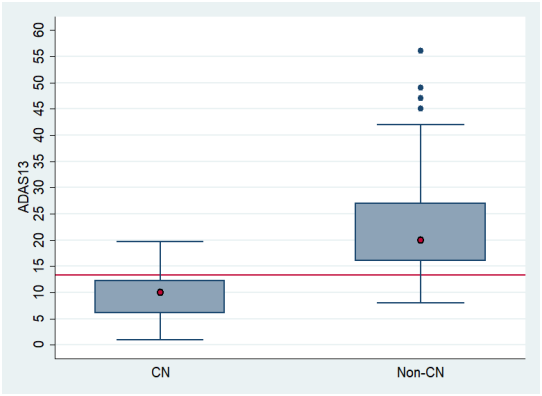
(b) ADAS4



(c) ADAS11



(d) ADAS13



(e) FAQ

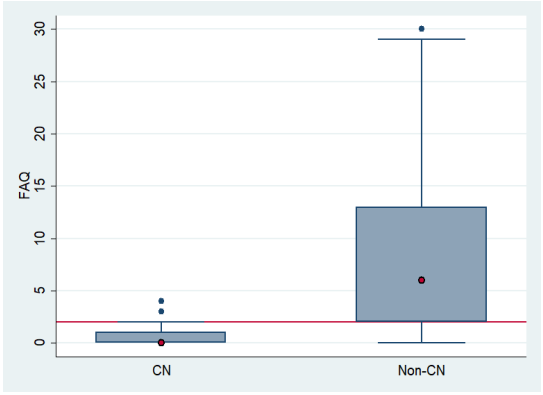
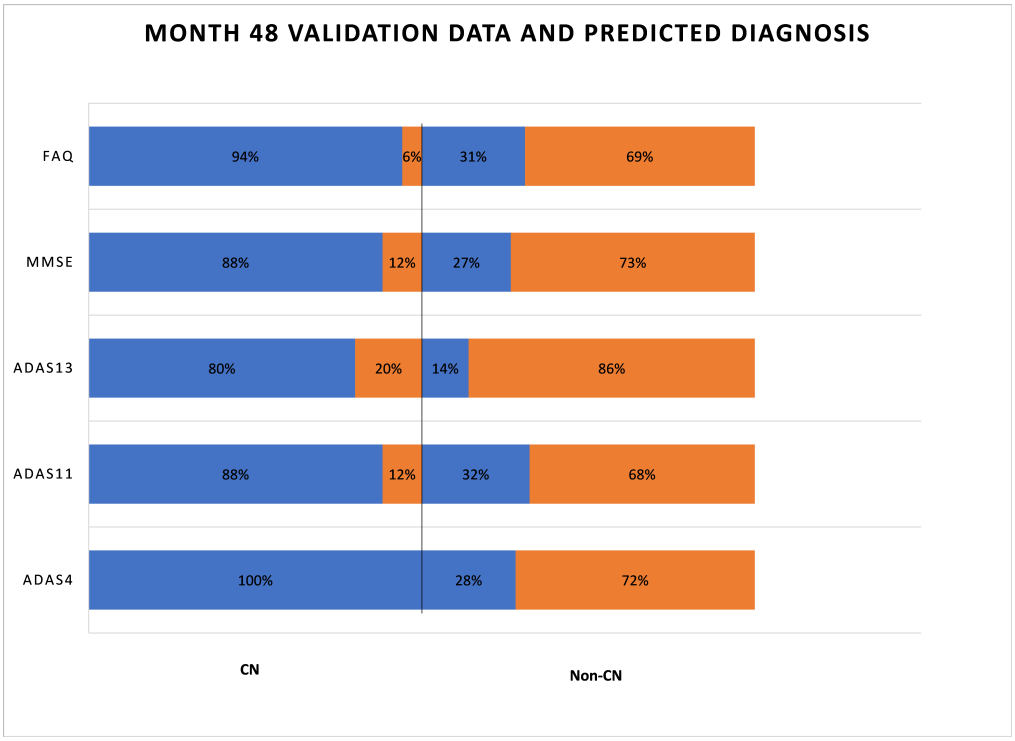
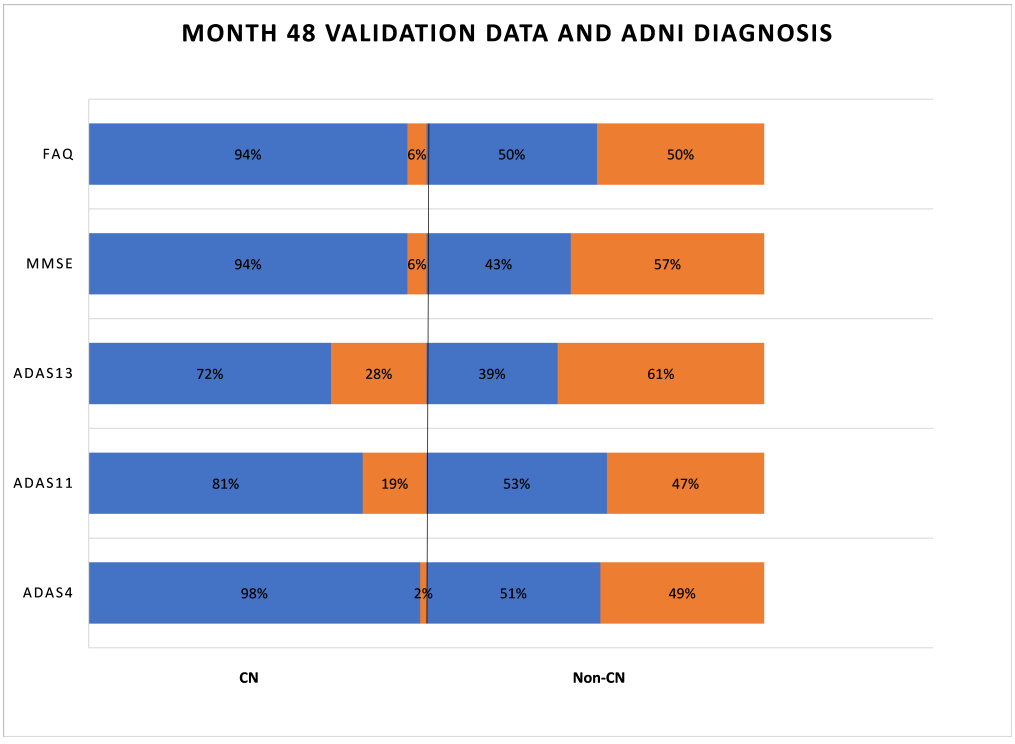


Figure 3: Proportion of Correct and Incorrect Diagnosis in CN and Non-CN Categories for Validation Data at Month 48



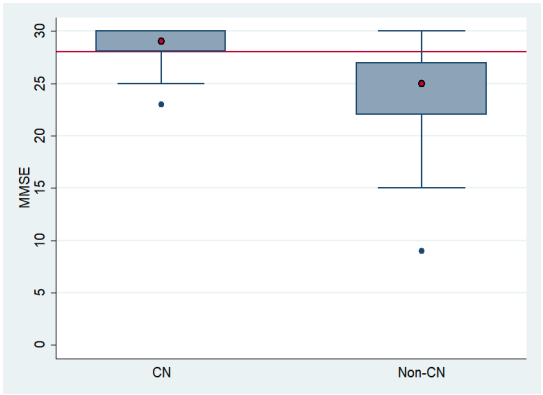
(3a)



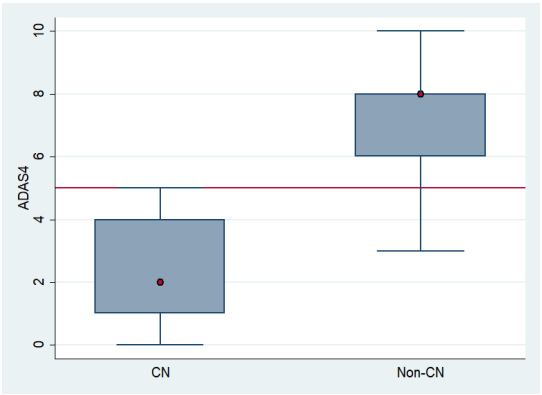
(3b)

Figure 4: Ranges of Validation Data Split by the Predicted Diagnosis (CN or Non-CN) Assigned by the Diagnosis Model at Month 72 (with Normal Cutoffs)

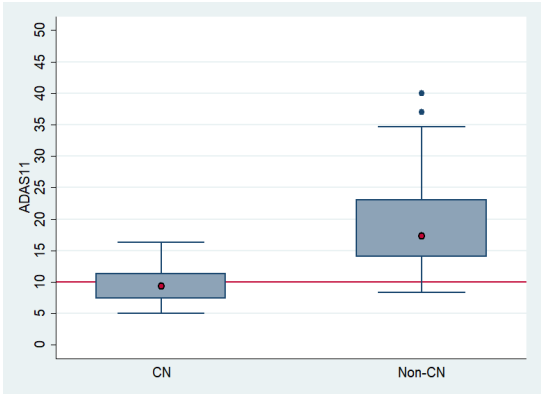
(a) MMSE



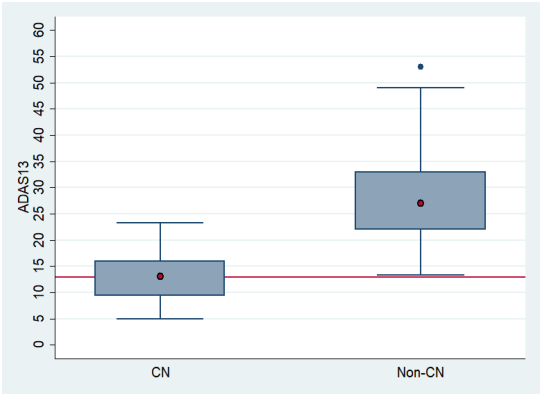
(b) ADAS4



(c) ADAS11



(d) ADAS13



(e) FAQ

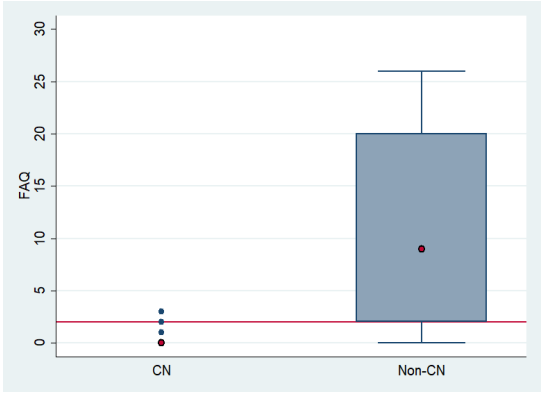
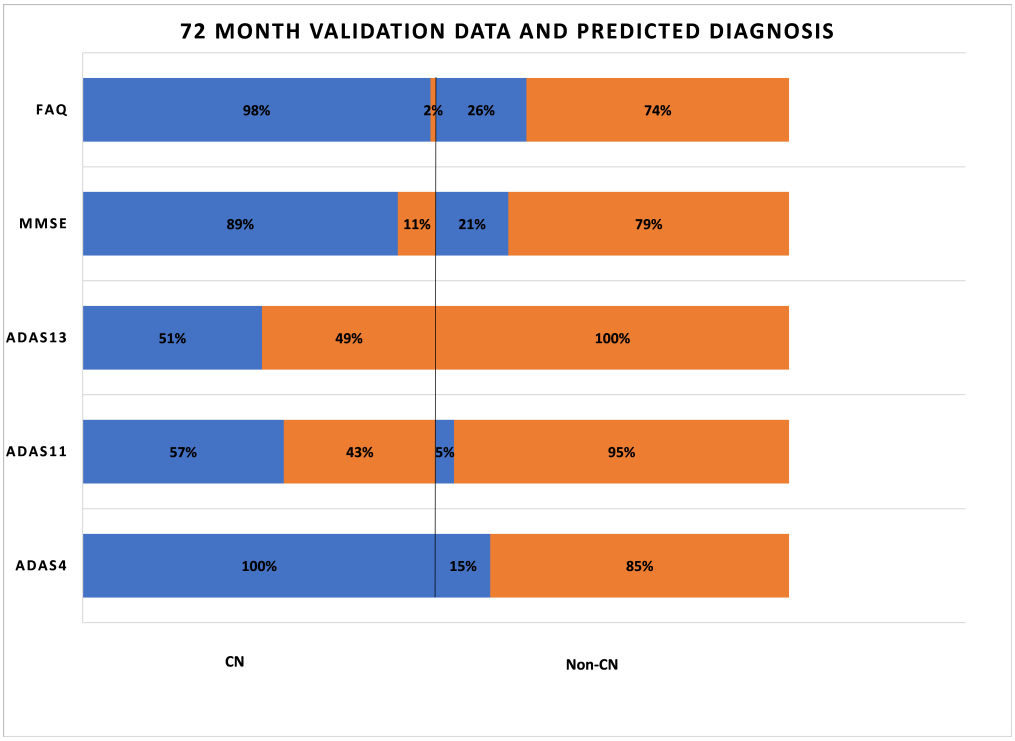
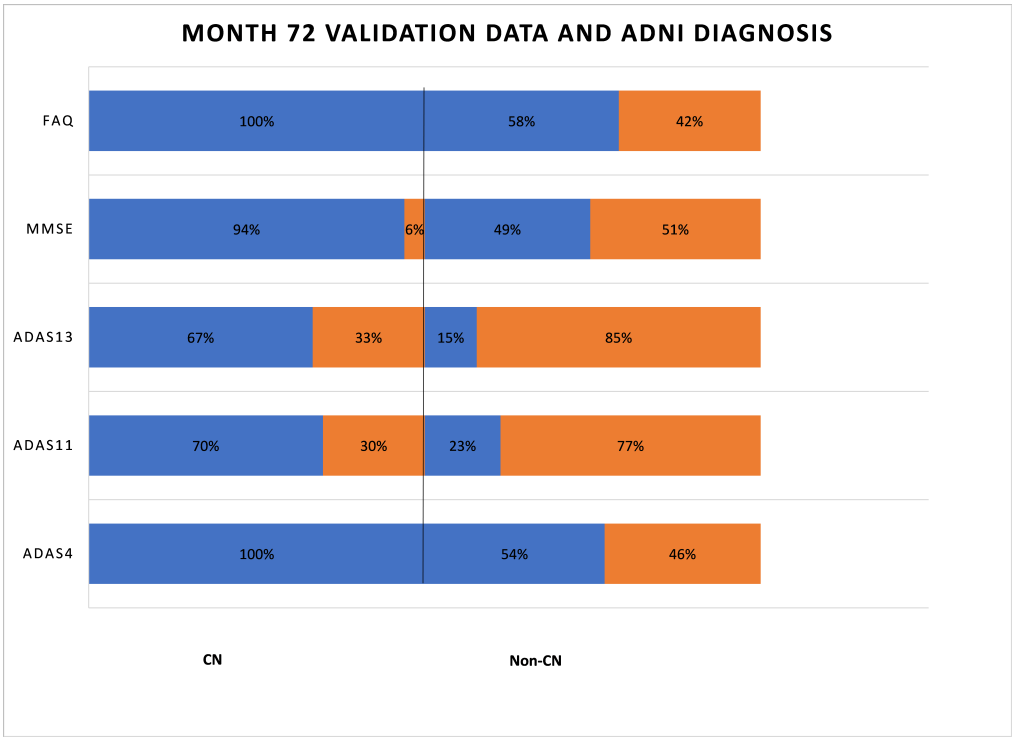


Figure 5: Proportion of Correct and Incorrect Diagnosis in CN and Non-CN Categories for Validation Data at Month 72



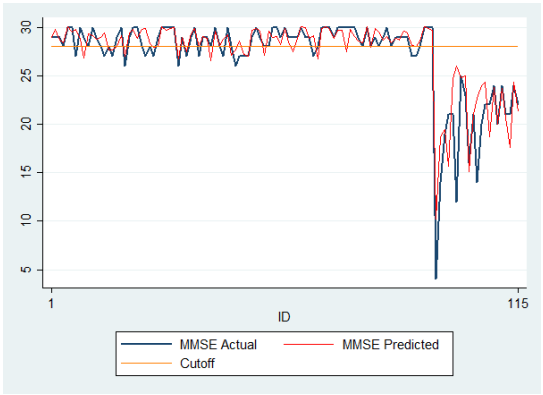
(5a)



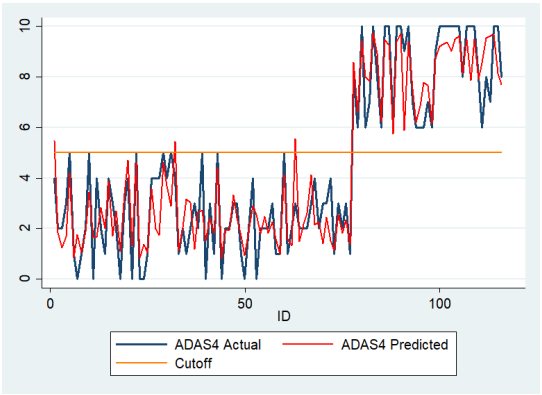
(5b)

Figure 6: Comparison of Actual and Predicted Values by Sequence Prediction Models at Month 48 for Validation Data (with Normal Cutoffs)

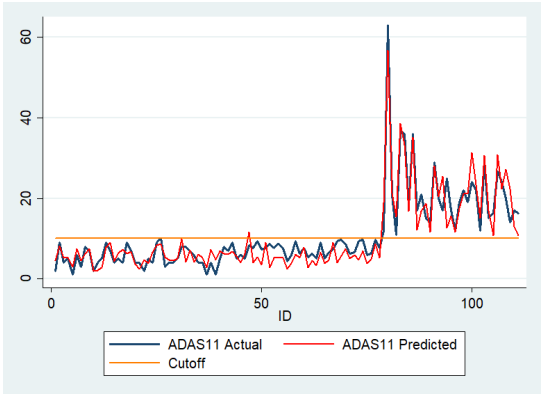
(a) MMSE Actual vs Predicted



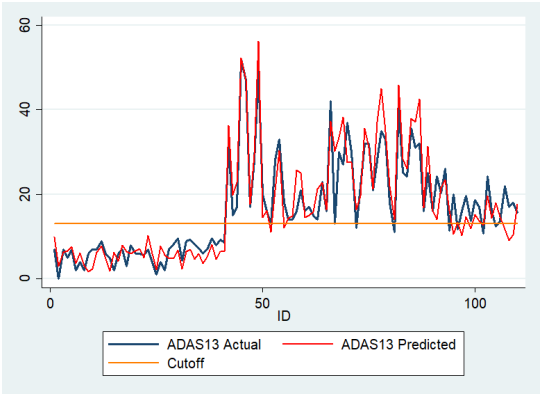
(b) ADAS4 Actual vs Predicted



(c) ADAS11 Actual vs Predicted



(d) ADAS13 Actual vs Predicted



(e) FAQ Actual vs Predicted

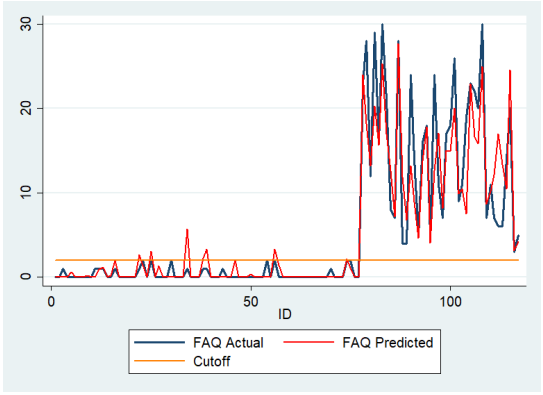
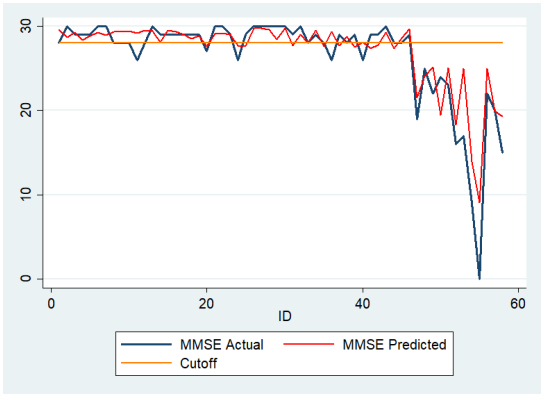
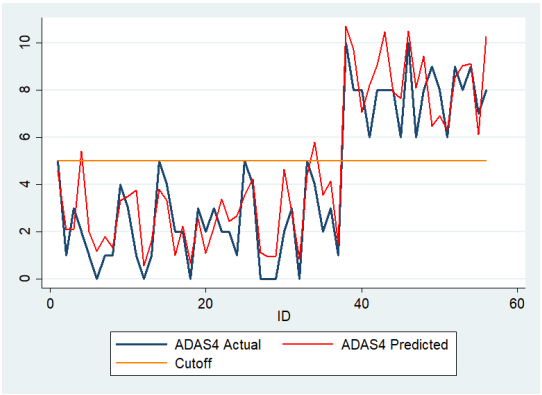


Figure 7: Comparison of Actual and Predicted Values by Sequence Prediction Models at Month 72 for Validation Data (with Normal Cutoffs)

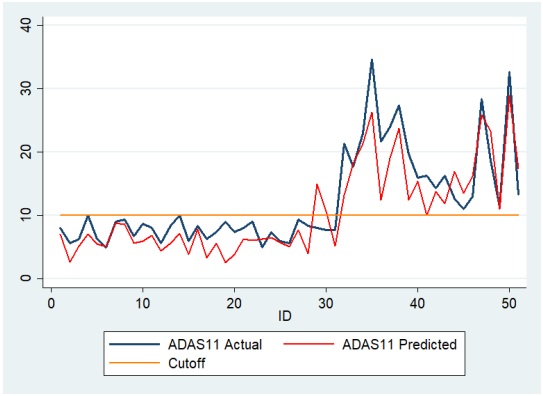
(a) MMSE Actual vs Predicted



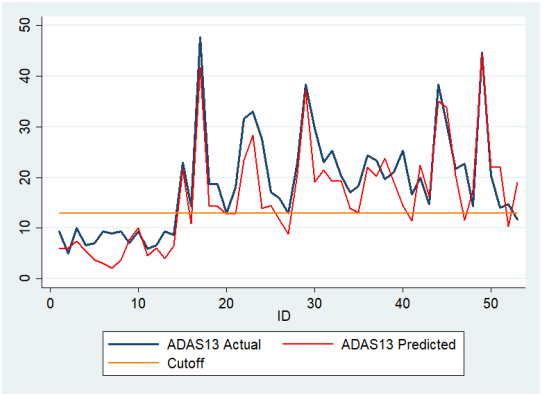
(b) ADAS4 Actual vs Predicted



(c) ADAS11 Actual vs Predicted



(d) ADAS13 Actual vs Predicted



(e) FAQ Actual vs Predicted

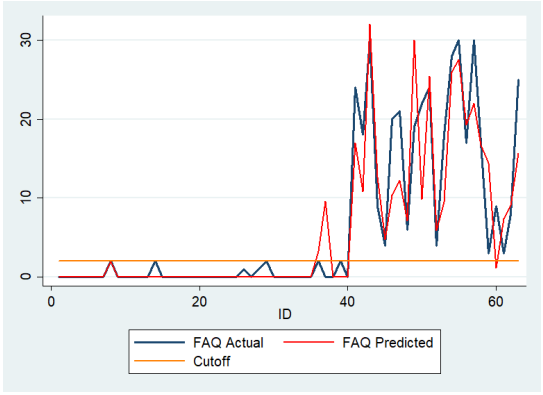
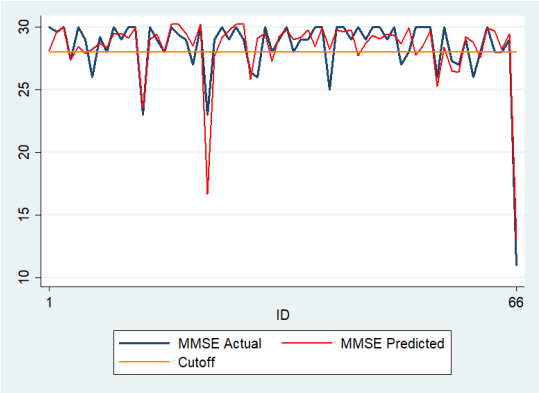
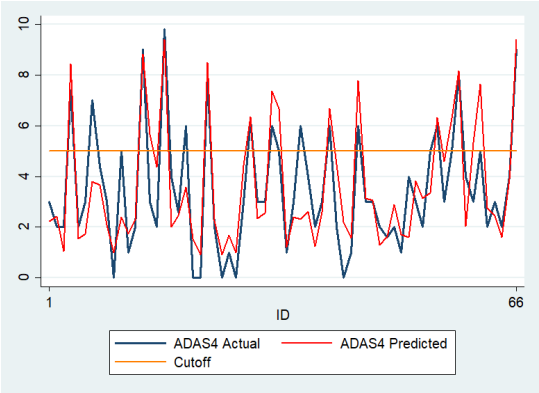


Figure 8: Comparison of Actual and Predicted Values by Sequence Prediction Models at Month 48 for 66 Test Dataset Subjects (with Normal Cutoffs)

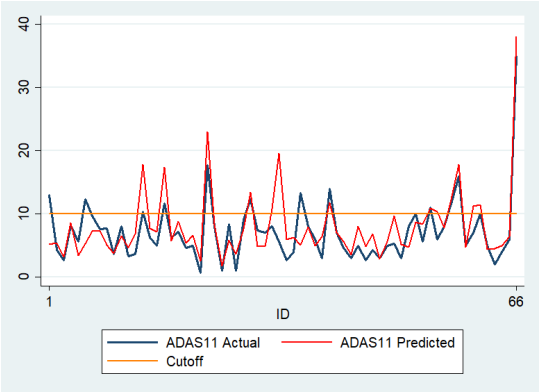
(a) MMSE Actual vs Predicted



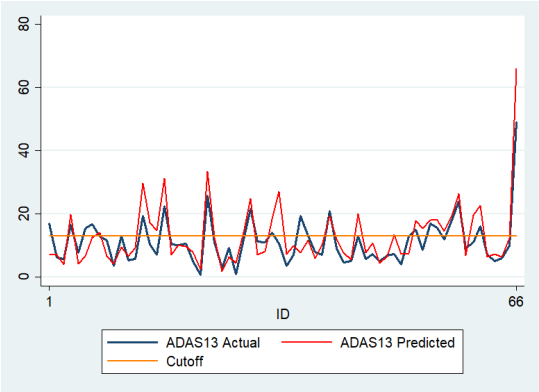
(b) ADAS4 Actual vs Predicted



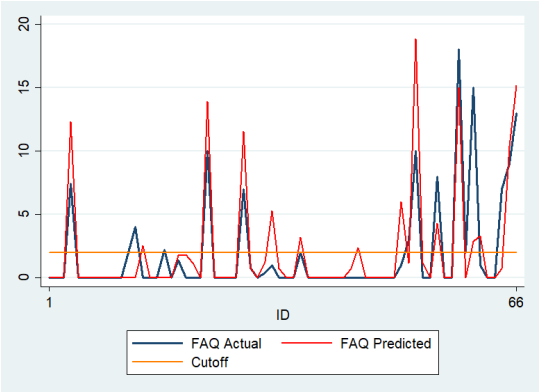
(c) ADAS11 Actual vs Predicted



(d) ADAS13 Actual vs Predicted



(e) FAQ Actual vs Predicted

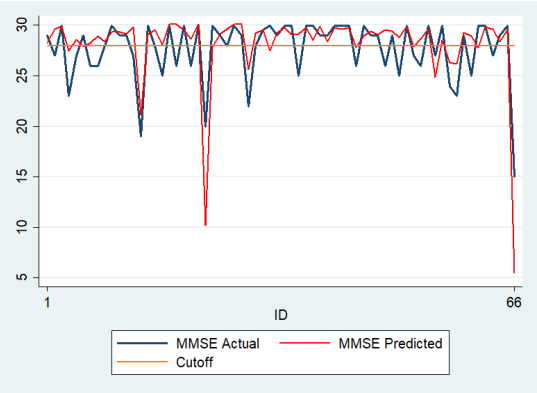


(f) % of Scores on the Same Side of the Cutoff Line for Month 48

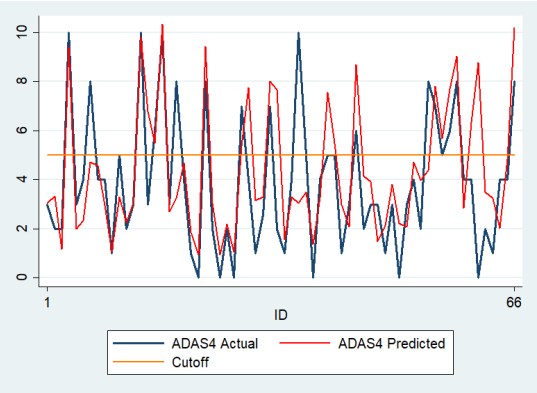
Test	% Same Side of Cutoff
MMSE	80.0%
ADAS4	91.0%
ADAS11	86.0%
ADAS13	76.0%
FAQ	88.0%

Figure 9: Comparison of Actual and Predicted Values by Sequence Prediction Models at Month 72 for 66 Test Dataset Subjects (with Normal Cutoffs)

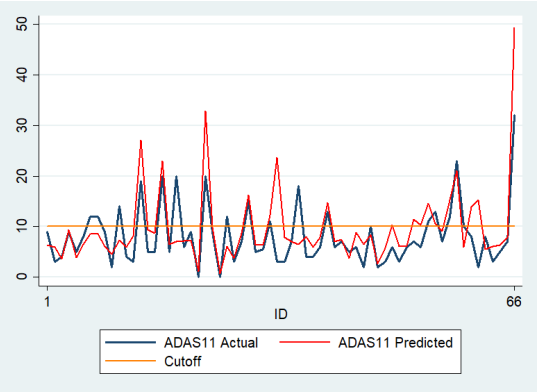
(a) MMSE Actual vs Predicted



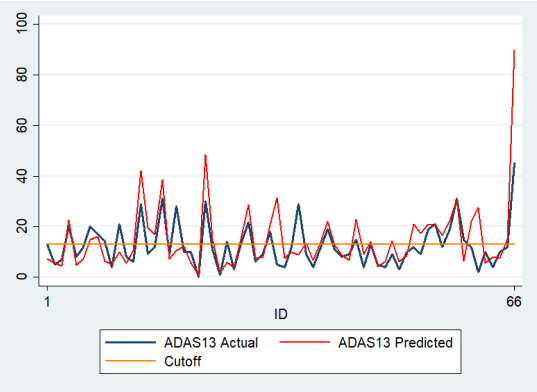
(b) ADAS4 Actual vs Predicted



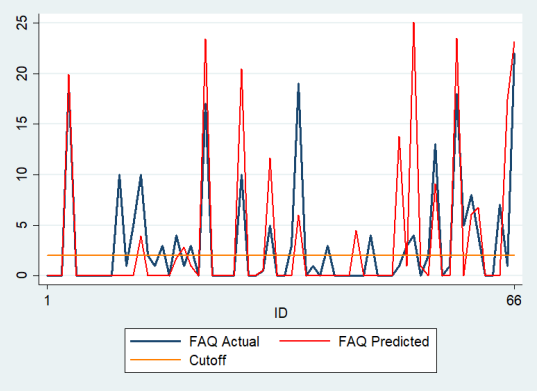
(c) ADAS11 Actual vs Predicted



(d) ADAS13 Actual vs Predicted



(e) FAQ Actual vs Predicted

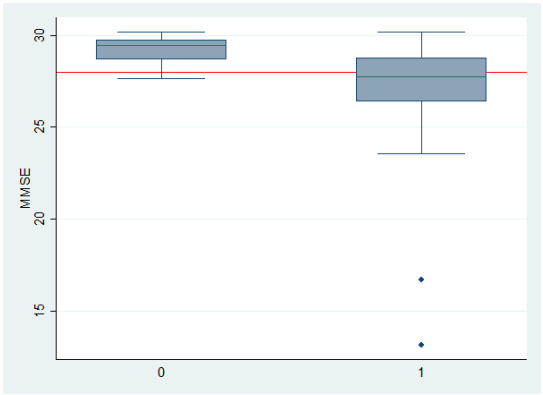


(f) % of Scores on the Same Side of the Cutoff Line for Month 72

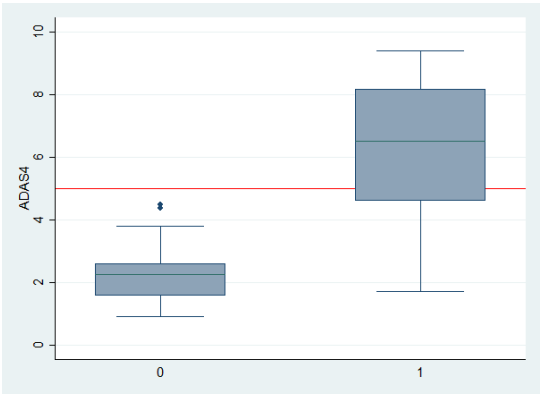
Test	% Same Side of Cutoff
MMSE	73.0%
ADAS4	82.0%
ADAS11	83.0%
ADAS13	76.0%
FAQ	79.0%

Figure 10: Ranges of Test Data Split by the Predicted Diagnosis (CN or Non-CN) Assigned by the Diagnosis Model at Month 48 (with Normal Cutoffs)

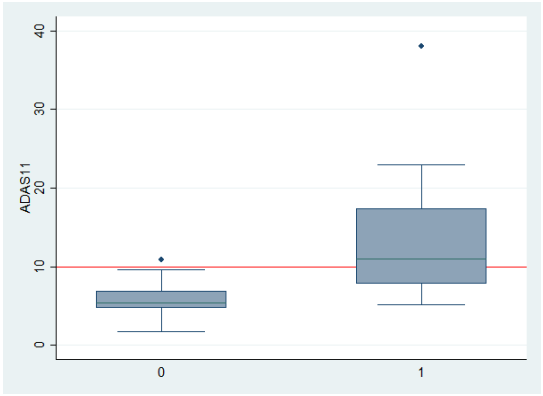
(a) MMSE



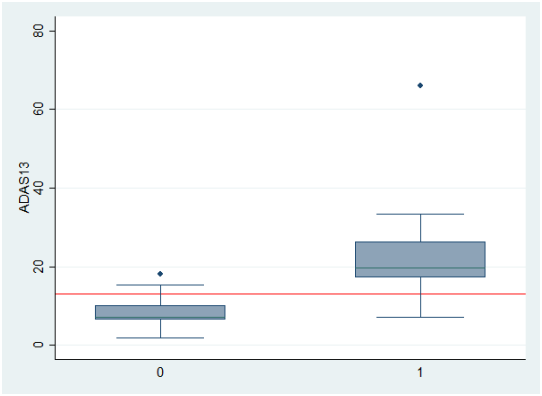
(b) ADAS4



(c) ADAS11



(d) ADAS13



(e) FAQ

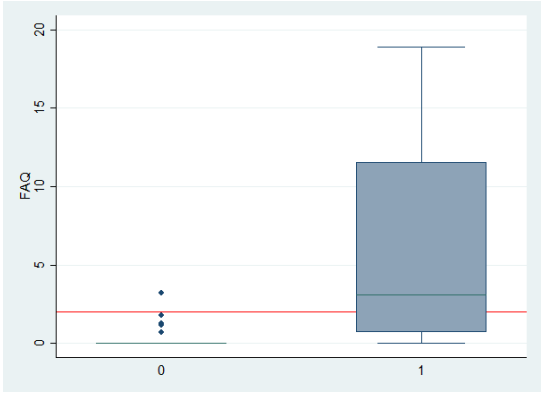
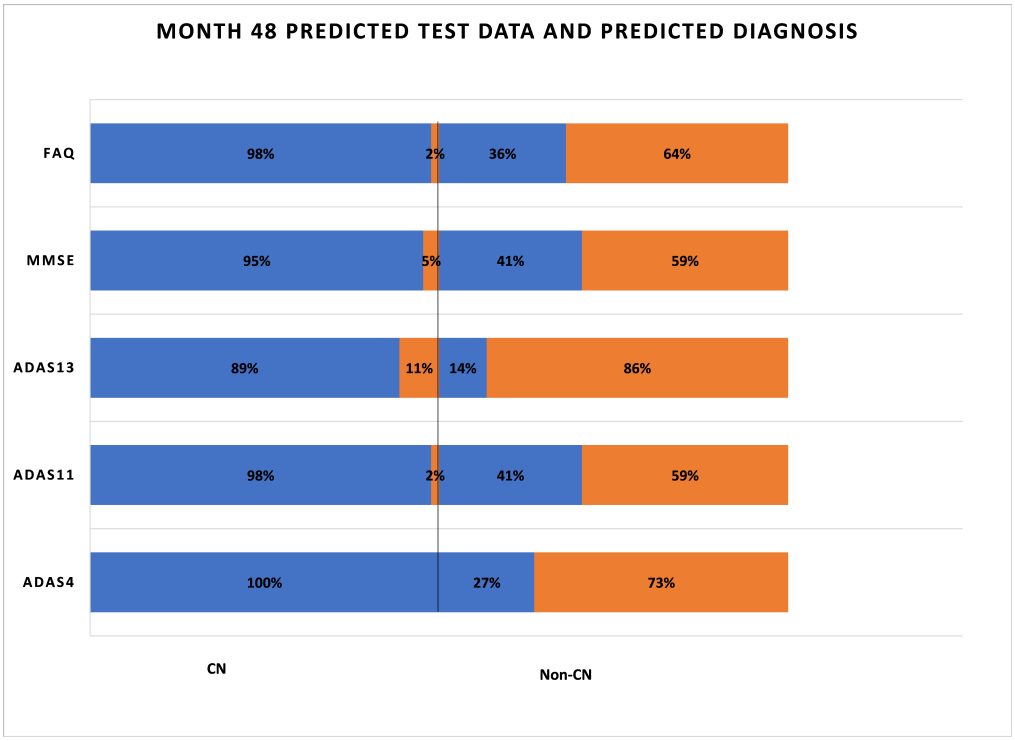
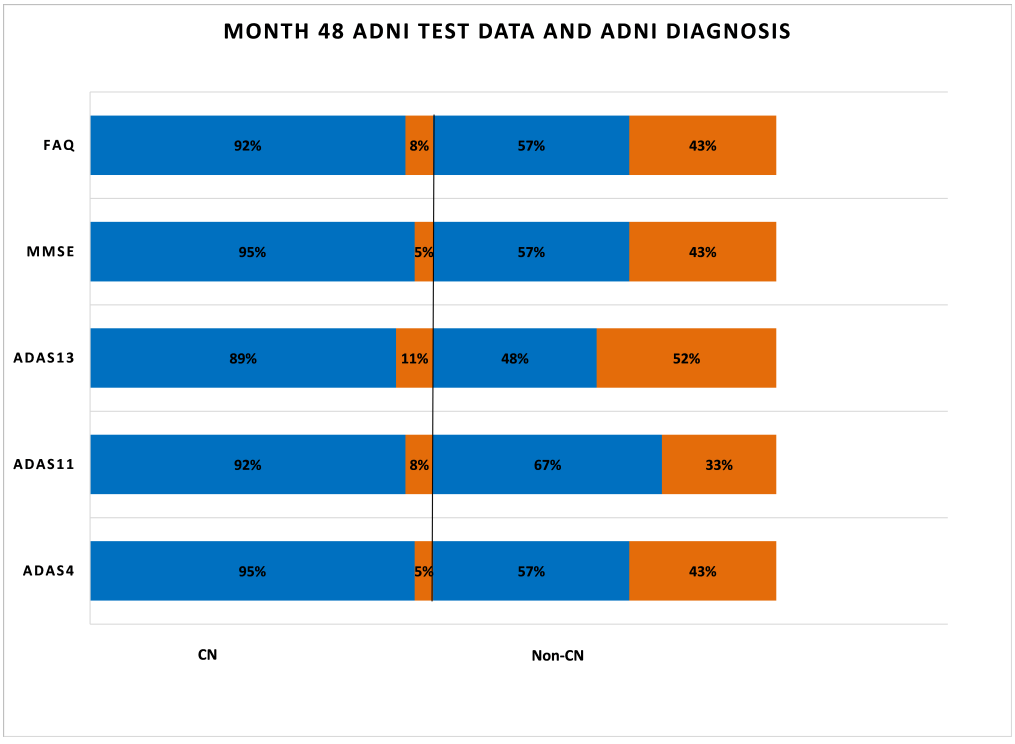


Figure 11: Proportion of Correct and Incorrect Diagnosis in CN and Non-CN Categories for Test Data and Predicted Test Data at Month 48



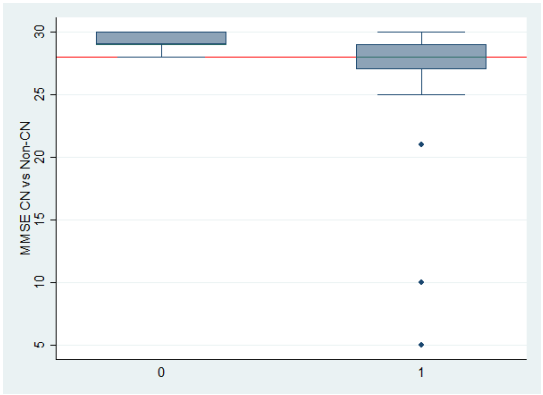
(11a)



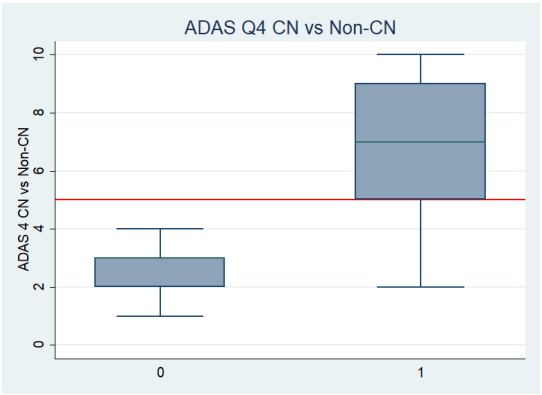
(11b)

Figure 12: Ranges of Test Data Split by the Predicted Diagnosis (CN or Non-CN) Assigned by the Diagnosis Model at Month 72 (with Normal Cutoffs)

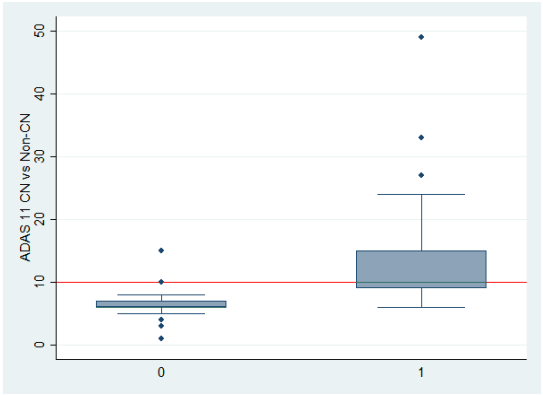
(a) MMSE



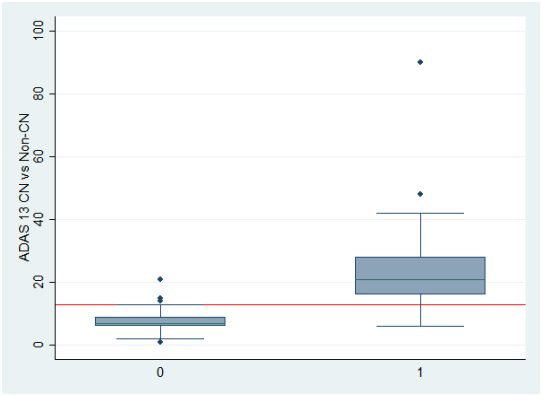
(b) ADAS4



(c) ADAS11



(d) ADAS13



(e) FAQ

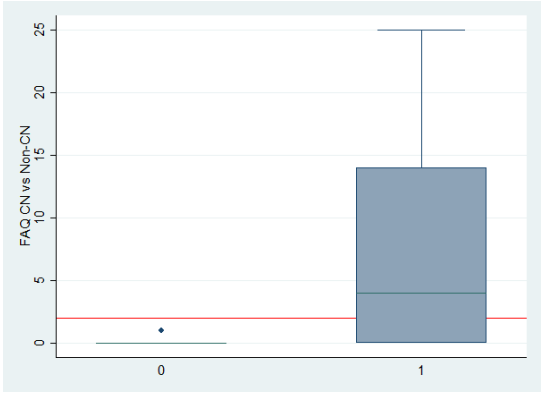
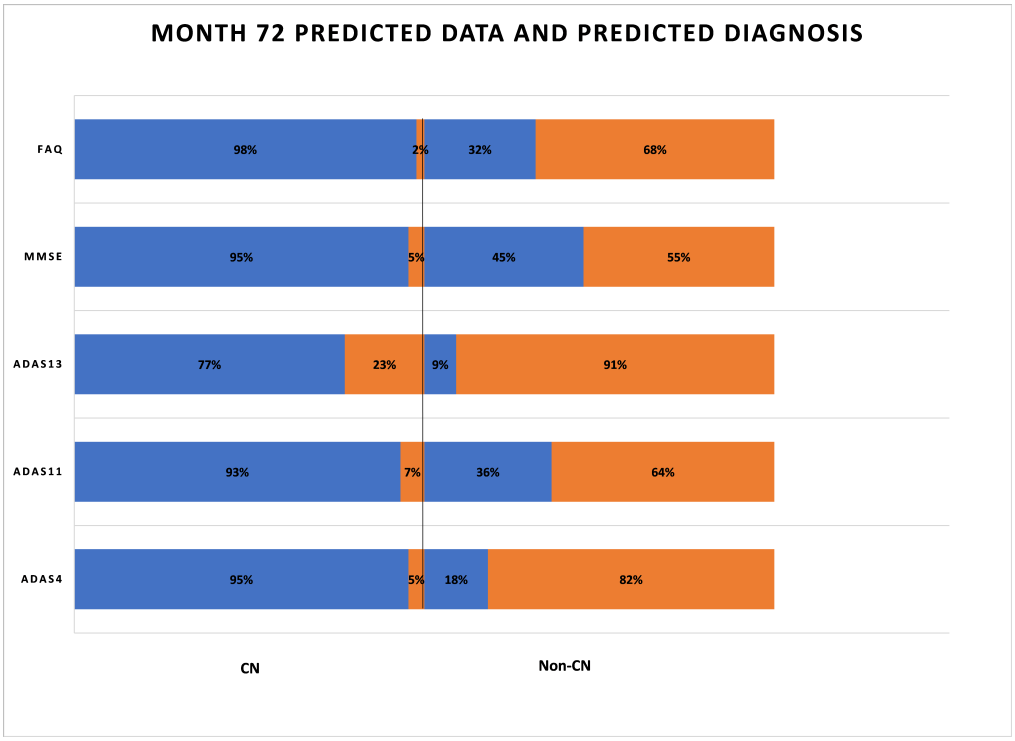
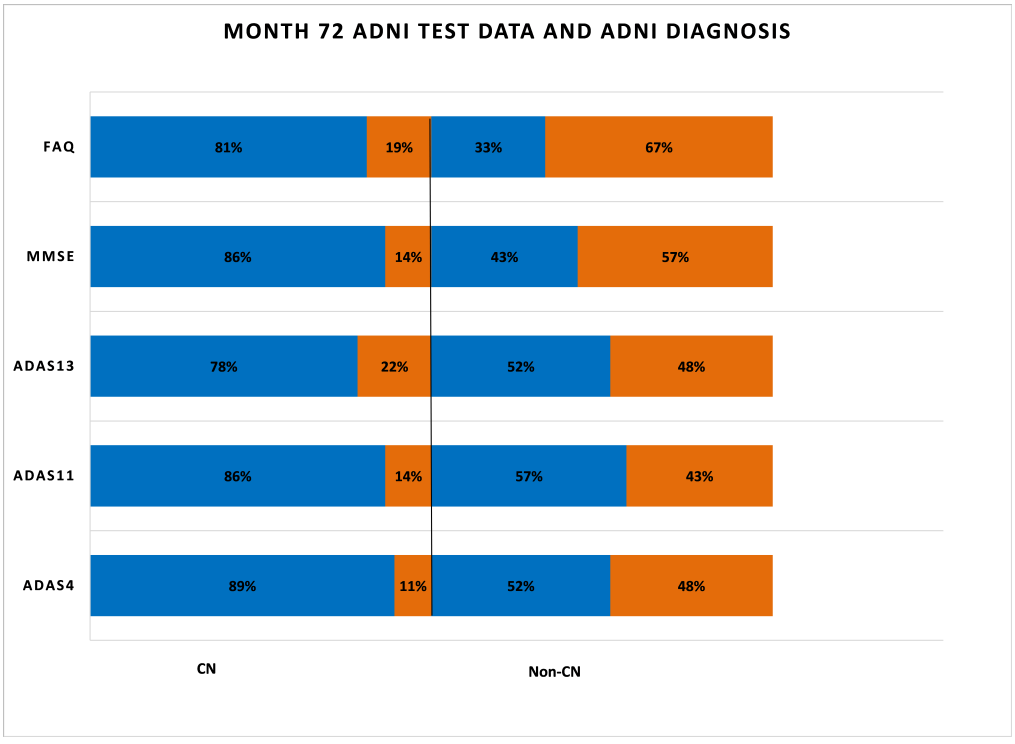


Figure 13: Proportion of Correct and Incorrect Diagnosis in CN and Non-CN Categories for Test Data and Predicted Test Data at Month 72



(13a)



(13b)