RESEARCH

The correlation of everyday cognition test scores and the progression of Alzheimer's disease: a data analytics study

Fadi Thabtah[*] ![ORCID], Robinson Spencer and Yongsheng Ye

Abstract

The process of diagnosing dementia conditions, especially Alzheimer's disease, and the cognitive tests that are involved in this process, are important areas of study. Everyday Cognition (ECog) is one test that can be used as part of Alzheimer's disease diagnosis to measure cognitive decline in different areas. In this study, we investigate two versions of the ECog test: the study partner reported version (ECogSP), and the patient reported version (ECogPT). We compare these, using statistical analysis and machine learning techniques, to create classification models to demonstrate the progression in ECog scores over time by using the Alzheimer's Disease Neuroimaging Initiative longitudinal data repository (ADNI); participants are classed with having normal cognition, mild cognitive impairment, or Alzheimer's disease. We found that participants who are diagnosed with Alzheimer's disease at baseline, or during a subsequent visit, tend to self-report consistent ECogPT scores over time indicating no change in cognitive ability. However, study partners tend to report higher and increasing ECogSP scores on behalf of participants in the same diagnosis category; this would indicate a degradation in the participant's cognitive ability over time, consistent with the progress of Alzheimer's disease.

Keywords: ADNI, Alzheimer's disease, Cognitive tests, Data analytics, Dementia, Everyday cognition, Longitudinal study, Machine learning

Introduction

Dementia is a progressive condition that affects the cognitive function of approximately 50 million people worldwide, with almost 10 million new cases being diagnosed every year [38], making it a key focus of healthcare institutions globally. Currently there is no cure for dementia or its leading condition—Alzheimer's disease (AD), but when diagnosed early there are intervention management plans available that can help patients cope with some of the symptoms [23]. Consequently, the detection and diagnosis of AD is an important objective for those studying neurodegenerative conditions.

A vast number of medical assessments have been created or adapted to detect AD, including biological tests such as positron emission tomography (PET), magnetic resonance imaging (MRI) scans, and cerebrospinal fluid (CFS) measurements taken from a lumbar puncture. However, most commonly, the diagnosis of AD is made by clinicians according to various cognitive tests. These are memory or cognitive ability-based questionnaires that are taken by the patient or a study partner such as caregiver.

Several research studies, using AD-based data, have studied the relationship between different combinations of medical tests as well as their diagnostic performance [3, 5]. These Studies have contributed to a better understanding of AD and the building of an effective and comprehensive system for its diagnosis. Many different cognitive tests have been compared to a wide variety of covariates, different biomarkers, and risk factors [10, 13, 33]. Additionally, different modelling approaches and data analytical methods have been used to identify the relationship between these tests and measure the effectiveness of each test at modelling the progression of AD

*Correspondence: fadi.fayez@manukau.ac.nz
Digital Technologies, Manukau Institute of Technology, Auckland, New Zealand

---

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 2 of 11

[19, 24]. Still, there are many different approaches yet to be attempted and more information to be uncovered about the relationships between these cognitive tests and AD progression.

This research investigates the longitudinal relationship between two versions of the commonly used Everyday Cognition (ECog) test, a cognitive test that was found to be comprehensive and responsive to change over time [14]. The ECog test is designed to measure cognitive decline in six different domains and is typically answered on behalf of the patient by a knowledgeable informant—we will refer to this version of the test as the Everyday Cognition Study Partner (ECogSP) test. Alternatively, the questions can be answered directly by the patient—we will refer this to as the Everyday Cognition Patient (ECogPT) version. However, previous studies have demonstrated that self-reported cognitive decline from a patient with AD is often less accurate than their study partner's-reported cognitive decline assessment [7, 29].

While ECog was originally designed as an informant-based test to overcome the well-known problems with self-reported cognitive decline, we could not find any studies that use in-depth correlation statistical analysis on a longitudinal dataset to examine the difference between test versions. Therefore, the aim of this study is to compare the patient reported ECogPT score to the study partner reported ECogSP score for the purpose of detecting cognitive decline over time. This analysis therefore re-iterates the differences between Informant-reported and self-reported cognitive testing and specifically shows how ECogSP and ECogPT scores differ overtime for subjects in different diagnostic groups.

We used participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data repository with mild cognitive impairment (MCI) and AD [25]. The methodology is based on data analytics, in particular correlation, statistical, descriptive analyses and supervised learning. The research question we try to answer is: "How the different versions of the Everyday Cognition test scores indicate the progression of dementia over time for patient with different baseline diagnoses?" We have created a number of visualisations to demonstrate the changes in ECogPT and ECogSP scores over-time for patients grouped by different baseline diagnoses. Repeated incremental pruning to produce error reduction (RIPPER) [9], partial C4.5 decision list (PART) [16], C4.5 decision tree (C4.5) [28], and Random Forest [6] machine learning algorithms were used to build models to predict if a patient diagnosed with MCI at baseline will progress to AD within 48 months based on ECog scores and demographic features. To overcome the issue of class imbalance and improve model performance, the

Synthetic Minority Oversampling Technique (SMOTE) [8] was used to balance the dataset.

"Literature review" section summarises a selection of research works related to longitudinal data analysis using data from the ADNI. "Dataset description" section presents the dataset, features, and the pre-processing done. "Methodology" section covers the methodology followed and "Data analysis" and "Machine learning results" sections show the results analysis and findings using machine learning models. In the last section we provide our conclusions.

Literature review

Moradi et al. [24] created a model to predict the scores of the Rey's Auditory Verbal Learning Test (RAVLT) [30, 32] based on grey matter density features derived from MRI scans in the ADNI dataset. The authors removed all observations with missing RAVLT scores and several observations that had outlier scores. Elastic net linear regression (enlr) [39] was then used to model RAVLT immediate and RAVLT percent-forgetting scores using whole brain grey matter density maps; these consisted of 29,852 features for each participant. The models were evaluated using 100 runs of tenfold cross validation [27]. Across all runs, the averages of the correlation score (R), coefficient of determination (Q^2), and mean absolute error (MAE) were found to be: R=0.50, $Q^2=0.25$, MAE=7.86 for RAVLT immediate, and R=0.43, $Q^2=0.185$, MAE=25.53 for RAVLT percent-forgetting. The author also considered data subsets that included only AD, or only MCI participants, as well as different combinations of the three categories. Interestingly, it was found that removing the MCI participants improved the performance of the model.

Ito et al. [19] used a mixed effect model [4] to predict Functional Assessment Questionnaire (FAQ) scores [26]. Again, the ADNI's longitudinal data was used, but this time features such as MMSC and CDR-SB scores, disease state, age, ApoE4 genotype, sex, and MRI-based biomarkers such as hippocampal volume, were included. The authors justified using a linear model to predict FAQ scores as they were only looking at three years' of data in the ADNI1 dataset and so expected the progression of FAQ scores to be relatively linear in this timespan. The authors noted that the distributions of FAQ scores, particularly around the end points (0 and 30), were not normal and would cause problems when fitting their model. Consequently, the standard approach was compared to a censored method where FAQ scores would be scaled and transformed. The models were evaluated by simulating 500 datasets based on the original.

Davatzikos et al. [10] studied the longitudinal trend in MRI and cerebrospinal fluid CSF biomarkers in two

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 3 of 11

groups: MCI participants who in a follow-up visit converted to AD (MCI-C), and MCI participants who did not convert (MCI-NC) within the ADNI dataset. The authors used a marker derived from MRI data named SPARE-AD which represents a pattern of brain atrophy that has been linked with AD [11]. SPARE-AD ranges between 1 and -1, with higher values representing more severe brain atrophy. In this study, MCI-C participants had an average SPARE-AD score of 0.65 ± 0.44, and MCI-NC had an average of 0.22 ± 0.74. Using the WEKA platform [18], the authors built linear support vector machine-based models to distinguish between MCI-C and MCI-NC participants based on the SPARE-AD score as well as the SPARE-AD and CSF biomarkers: Aβ42 and t-tau. The models were evaluated using fivefold cross validation. The SPARE-AD only was the only model that achieved a sensitivity towards MCI-C of 94.7% and specificity of 37.8%. The model that also included Aβ42 and t-tau achieved a sensitivity of 84.2% and a specificity of 50%. It should be noted that the data of 239 patients with follow up visits was used for the SPARE-AD model; whereas the data for 120 patients only was used in the SPARE-AD and CSV model since participants with missing Aβ42 and t-tau values were removed.

Evans et al. [13] used data from the ADNI1 study to compare the changes in whole brain and ventricular volume over time in AD, MCI, and CN participants. Data was limited to follow up visits within 12 months of the initial visit. A one-way ANCOVA test was used (with p values < 0.0005) to determine that there was a significant difference in brain atrophy rates between participant groups when accounting for age, gender, and baseline volume. The authors also considered the cognitive tests ADAS-Cog [31] and MMSE [15] to determine if they could be associated with changes in whole brain and ventricular volume. They concluded there was strong evidence of an association between change in MMSE scores and change in whole brain volume in AD (p = 0.002) and MCI (p < 0.0005) participants. There was weaker but still significant evidence of an association between whole brain volume and ADAS-Cog changes in AD (p = 0.06) and MCI (p = 0.0001) participants.

Schuff et al. [32] applied a linear, mixed-effect model on the selected data extracted from ADNI, using hippocampus volume degradation (HVD) or ADAS test as the dependent variable to regress on time (independent variable). Since past measurements determine the current measurement, the authors have also applied the Markov Chain model and adopted a paired model to investigate if additional independent variable(s), i.e. age, diagnosis, ApoE4 (gene), cerebrospinal fluid (CSF) biomarkers etc., would increase the explanatory power. The

dataset selected consists of 12 months of longitudinal data on 112 CN, 226 MCI, and 96 AD participants, with a minimum of three MRI scans. The major findings of the study were:

(1) The protocols adopted in ADNI for MRI measurements are effective for regulating and minimising site-to-site variation, ensuring a high level of uniformity and data integrity

(2) MRI volumetric changes can be tracked over a short period of time, i.e. 6–12 months, and show that both MCI and AD participants have significant HVD over 6 months and this gradually increased over a 12 month period

(3) Faster HVD is associated with the ApoE4 (gene) in AD participants and lower CSF b-amyloid (Aβ1–42) in MCI participants respectively. This also proves hippocampus volumetric measures to be an effective biomarker for tracking AD progression.

Vemuri et al. [36] selected 71 AD, 19 aMCI, and 92 CN participants from ADNI. The study compared the annual changes of MRI and CSF biomarkers to assess the discrimination power among 3 groups (CN, aMCI, and AD) and if any correlation existed between the cognitive score changes and the selected biomarkers, and the relevance of the Apoe gene to the biomarkers' annual changes. The authors chose Aβ1–42 and t-tau as the CSF biomarkers, and ventricular (VC) as the MRI biomarker instead of entorhinal cortex (EC) or hippocampus (HC) as used in other studies [12]. After plotting the annual changes of those biomarkers among the participant groups, the authors have drawn the conclusion that VC changes are a significant (p < 0.001) intergroup, but the CSF biomarkers are not. The study has also concluded that VC changes are significantly linked to cognitive score changes in aMCI and AD participants, The ApoE4 (gene) does play a role in the VC changes, but not for the Aβ1–42 and t-tau CSF biomarkers. However, other studies have suggested that abnormal changes occurred to Aβ1–42, then t-tau, and lastly MRI during the progression of AD [20, 21]. The contradictory outcome of this study prompted the authors to consider the possibility of the CSF biomarkers of the selected sample becoming abnormal before any clinical symptoms surfaced.

Varon et al. [35] selected 50 AD, 89 MCI, and 49 CN participants, aged 55–81, from the ADNI1 dataset to ascertain the clinical application of visual ratings (VR) and volumetric measures (VM) of medial temporal degeneration (MTD) in dementia. The authors selected entorhinal cortex (EC) and hippocampus (HC), and two medial temporal lobe structures (MTLS) to investigate after researching relevant past studies; these established

Thabtah *et al. Health Inf Sci Syst* (2020) 8:24

Page 4 of 11

a high degree of correlation of the MTLS in the clinical cases of MCI and AD [12]. A 4-point scale of 0–2 was adopted for VR, where 0–0.5 indicated minimal degeneration, 0.5–1 indicated mild degeneration, 1–2 indicated moderate degeneration, and over 2 indicated severe degeneration, for HC and EC in this research. There are several notable findings after analysis:

(1) VM was better than VR-HC regarding CN and MCI discrimination, however, both VM and VR (HC & EC) had similar power in CN and AD discrimination at baseline

(2) VM (HC & EC) and VR (HC & EC) were both significant predictors in regard to MCI progressing to AD, but VR-EC surpassed VR-HC and VM-EC

(3) VM-EC is better than VM-HC for predicting MCI, which is a similar finding to other studies [2, 22]

(4) During the early stage of AD, HC's relevance can be dampened by other common medical conditions, i.e. depression, stroke, steroid intake, brain trauma, and seizures [17]

Dataset description

Since October 2004, with funding from private and public institutions, Dr Michael Weiner has led a team conducting a longitudinal study, ADNI, for the primary purpose of finding and validating biological markers to be used in AD clinical trials [1]. The study is ongoing after its inception more than a decade ago, having been through three different National Institute on Aging (NIA) funding cycles (ADNI 1, 2, and 3) and a bridging grant (ADNI GO); it is currently funded until 2022 [1]. Study participants were recruited from different medical and research facilities across the US and Canada, and also selected from a combination of normal aging elderly (NL), MCI participants, and those with confirmed AD [37]. These participants complete several cognitive tests and clinical tests when required, including DNA, blood test, urine test, CSF, MRI, and PET [25].

The dataset was obtained from the TADPOLE challenge, which was initiated by ADNI [34]. The combined dataset consists of 12,742 instances where each instance represents an individual visit to the clinic where measurements were taken. There was a total of 1907 features. Apart from the demographic features and those for clinical visit identification, four features related to the ECog cognitive test have been retained: the total and memory scores for both the patient-based version and study partner version of the test. A brief summary of the features selected with some basic statistics is given in Table 3 (Appendix).

The ECog medical test was developed by Farias et al. [14] to assess the cognitive degradation of the patient based on an interview with an informant such as the caregiver or a relative of the patient and asking questions related to the following six cognitive domains: memory, vocabulary and linguistic awareness, visual/spatial awareness, preparation, organisation, and divided attention. The test began with 138 possible questions, based on information provided by domain experts such as clinicians and neurologists, then reduced to a final 39 questions with four possible answers to each question:

1 = Better or no change compared to 10 years earlier.
2 = Questionable/occasionally worse.
3 = Consistently a little worse.
4 = Consistently much worse.

Each item within the test has an individual factor weighting and contributes to a total score, as well as a score for its related cognitive domain. Scoring is based on a scale of 1–4, with '1' being the most benign and '4' being most severe in terms of cognitive decline. In this study, we have focused on the ECog Total score which covers all cognitive domains and the ECog Memory score.

ECog was initially designed to be a study partner-(informant) based test, however, within ADNI, the participant also responds to the same questions in a patient-based test. The aim of this study is to compare the informant-based ECog test to the patient-based ECog test, hence this study examines four different ECog test scores: the patient-reported memory score (ECogPT Mem), the patient-reported total score (ECogPT Total), the study partner-reported memory score (ECogSP Mem), and the study partner-reported total score (ECogSP Total).

In ADNI dataset, the average ECog assessment scores provided by study partners are slightly higher than the participant test score, 2.14 vs 2.07 for the memory test
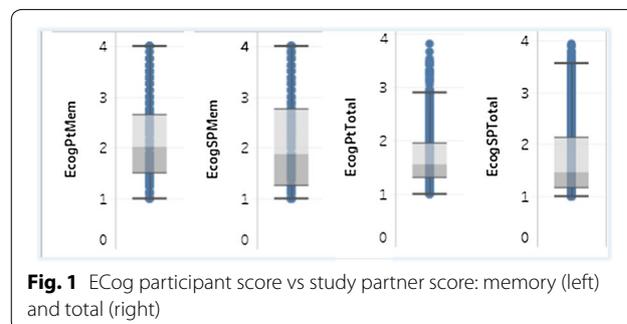
Fig. 1 ECog participant score vs study partner score: memory (left) and total (right)

Thabtah *et al. Health Inf Sci Syst* (2020) 8:24

Page 5 of 11

Fig. 2 Baseline diagnosis, gender and boxplot for age
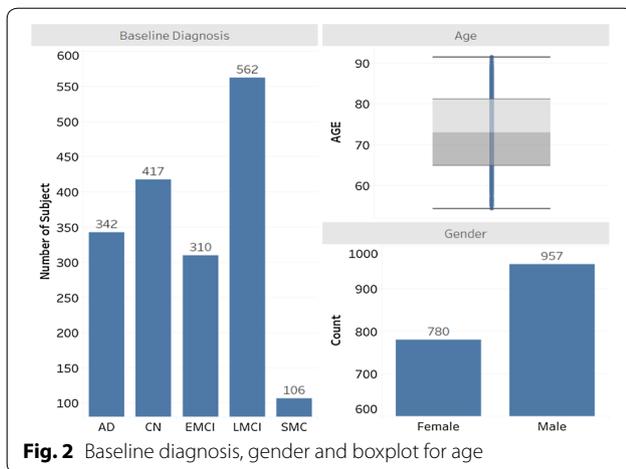
scale and 1.82 vs 1.68 for the total test scale, as seen in Fig. 1.

As seen in Fig. 2, the participants can be categorised into five groups in baseline clinical diagnosis: AD, EMCI (Early MCI), LMCI (Late MCI), SMC (Subject Memory Concern), and CN (Cognitive Normal). The gender ratio is 1.23: 1 with males exceeding females by approximately 23%. The average age of the participants was 73.8, ranging from 54.4 to 91.4. Participants averaged 7.3 visits to the clinical centres during the period of study. Their educational level ranged from 4 to 20 years with an average of 16 years. The majority of the participants classified their race as white (1605 out of 1737). In regard to marital status, 1311 participants were married, while the other 410 were widowed, divorced, or single.

Before pre-processing, the average ECogPT score was lower than the ECogSP score for both memory and total scores. This indicates the study partner considered the cognitive decline to be greater than what the participant self-rated. Interestingly when selecting for baseline diagnosis the same trend was retained for the AD group but was completely opposite for the NL group in which the participants estimated their cognitive decline to be greater than their study partner reported.

Compared to the other cognitive tests, ECog is fairly new, having been introduced in 2008. Hence, it was not included in ADNI 1 and ADNI GO as part of the clinical testing, however in ADNI 2, returning patients from 1 and GO were tested using ECog. This has resulted in some participants within the study having no ECog scores recorded during their baseline visit but during subsequent visits. To include such participants in our experiment, we adjusted all baseline visits to be the subject's first visit when ECog test scores were taken.

Any prior visits from these participants where ECog scores were missing have been removed and two new features have been created:

'Month_Adjusted' for recording the month of visit since first ECog test score was taken
'DX_bl_Adjusted' for recording the diagnosis made by clinicians at the adjusted baseline.

Due to a new baseline being introduced, some of the data requires further processing so that it is aligned with the new baseline. The missing values in 'EcogPtMem_bl', 'EcogPtTotal_bl', 'EcogSPMem_bl', and 'EcogSPTotal_bl' which record the bassline scores for each ECog version have been amended. Also, to simplify and align the diagnostic outcomes:

The original baseline diagnostic categories of AD, LMCI, EMCI, SMC and CN have been reclassified to AD, MCI (LMCI + EMCI) and NL(SMC + CN).
The final clinical diagnostic categories of AD, AD to MCI, MCI, MCI to AD, MCI to NL, NL, and NL to MCI have been reclassified as AD (AD, MCI to AD), MCI (AD to MCI, MCI, NL to MCI), and NL (MCI to NL, NL).

Another new feature, 'MCI_Change' has also been introduced to explore the correlation between the ECog score and the baseline MCI participants who have progressed to AD, and the participants who remain as MCI. This feature has three possible values: "Not MCI" if the participant is not diagnosed with MCI at baseline, "No Change" if the participant was diagnosed with MCI at baseline and diagnosed with MCI during their last visit within 48 months. Finally, "MCI to AD" if the participant was diagnosed with MCI at baseline but diagnosed with AD during their last visit within 48 months.

Considering the adjusted baseline, the number of returning visits declined over a period of 72 months. In fact, the number of visits dropped to only 32.6% of participants after 48 months, and 13.2% after 60 months. Therefore, it makes sense for this study to limit the time period to 48 months, otherwise the analysis outcome could be skewed significantly due to the participant drop off.

After the above pre-processing has been performed, only 1183 participants are left in the dataset, 544 female and 639 male, aged from 55–91, and the number of participant diagnosed as MCI is more than twice those with AD. the average age and education level of the participant among three diagnosis groups remained very similar trough pre-processing.

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 6 of 11

Fig. 3 Methodology graph

Fig. 4 Average ECogSP Total vs average ECogPT total scores overtime
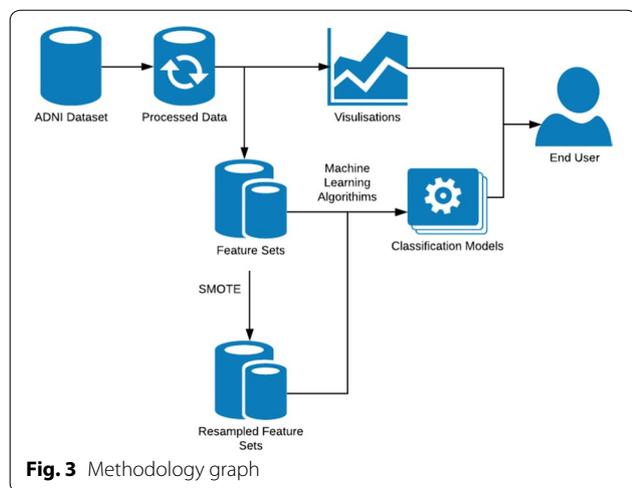
Methodology

We have followed the methodology demonstrated in Fig. 3 throughout our analysis. Firstly, we obtained the publicly available ADNI TADPOLE dataset [34] and undertook research and some descriptive analysis to gain an understanding of the domain and dataset. Shown in "Literature review" section, our research was focused on longitudinal analysis projects using the ADNI dataset as well as some background study on Alzheimer's disease and the ECog test. Descriptive analysis in shown above in "Dataset description" section and includes distributions of Diagnosis and ECog scores in addition to demographic information and dataset background.

Next, based on some issues we found in the dataset which are described in "Dataset description" section we underwent a significant pre-processing phase. This Included removing features not related to ECog scores, Diagnosis, and visit information, removing instances with missing ECog Scores including visits prior to the ECog test being introduced in ADNI, and limiting the instances we use to visits within 48 months of a subject's baseline visit since there is a significant drop of returning visits after that time period.

We performed some simple linear regression analysis and used the program: Tableau to create a selection of visualisations that demonstrate the difference in ECogSP and ECogPT scores overtime. Pearson's product-moment r, or r coefficient, was used to calculate the correlation between ECog test scores overtime. The absolute of coefficient r can lie between 0 and 1, where 0 indicates no correlation and 1 indicates a perfect linear association. Furthermore, a positive r shows that any addition in one variable leads to an increase of the other variable, while a negative r shows any subtraction in one variable leads to a decrease of the other variable.
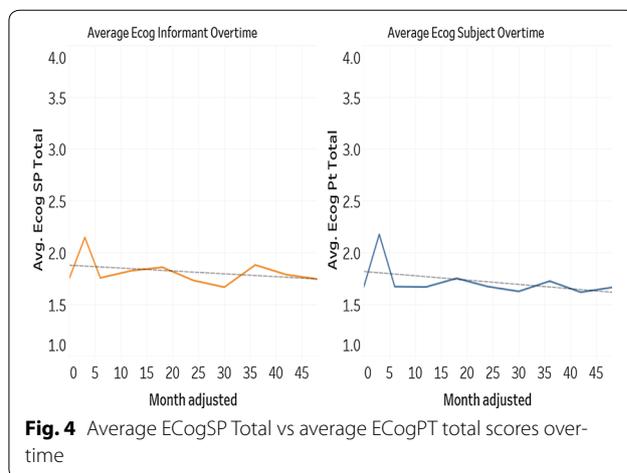
We employed different classification algorithms including RIPPER, PART, C4.5, and Random Forest on three feature sets: ECogSP scores with demographic information, ECogPT scores with demographic information, and only demographic information to assess each test version's ability to predict the progression of MCI subjects to AD from baseline test scores. tenfold cross validation was used to produce evaluation metrics for each model. Accuracy, precision, and recall were recorded for each model over 10 experimental runs and their averages reported in "Data analysis" section. A paired corrected t-test (with $p < 0.05$) was used to determine if differences in model performance across feature sets was statistically significant.

To reduce class imbalance in the data used to train classification models we employed the SMOTE method to create a resampled dataset to reduce any possible bias toward the majority class in the dataset. The machine learning experiment was then repeated on the new resampled data and the performance of models trained on the resampled dataset was compared to models trained using the original dataset.

Data analysis

Data visualisations demonstrate the trend in ECog Total scores over time for the three diagnostic groups: AD, MCI, and NL, and show the differences in trend for MCI patients who progress to AD, and MCI patients who remain with MCI diagnosis within 48 months. Correlation analysis is appropriate for this data since it can reveal the relationship between ECog score and the months since baseline visit, to determine if the ECog score changes over time. The analysis has been conducted in the Tableau Desktop 2019.4.0 platform and performed on an i7 desktop with 32 GB RAM.

Thabtah *et al. Health Inf Sci Syst* (2020) 8:24

Page 7 of 11

In Fig. 4, the average ECog scores between participants and study partners are plotted against each other side-by-side. The difference between these ECog scores isn't obvious with very similar gradient and intercept and clearly both have a downward trend. Both r coefficients are very similar, -0.351 for ECogSP and -0.429 for ECogPT, which means the study partner's average ECog score is more correlated to the months since baseline visit than the participant's average ECog score. This also indicates that the average ECog scores are reducing over the course of 48 months, which means the overall cognitive degradation has reversed and the participant's health is improving. The most likely reason behind this outcome is the effect of averaging, i.e. extreme outliers distorting the central tendency as visits drop off overtime.

To investigate further, instead of averaging all ECog scores regardless of the diagnosis, the average total ECog scores have been plotted against the visit month for three different diagnosis groups: AD, MCI, and NL (control group) as seen in Fig. 5a and b. Figure 5a shows the average ECogSPTotal score over time, the trends are distinctly different among the three groups. The r coefficients for AD, MCI, and NL are 0.777, 0.724, and 0.686 respectively. In respect to month of visit, the average ECogSP score for the AD group is more correlated than for the MCI and NL group. The AD group has the highest average ECogSP score which has increased drastically over time; this was

expected due to confirmed AD diagnosis. Both the MCI and NL group have a significantly lower average ECogSP score in comparison to the AD group.

In Fig. 5b, the trends are similar in the three groups, except for the AD group which has more fluctuation. An explanation for this is the decrease of visits in certain months. The average ECog score for all three groups is slightly trending down indicating that the participant's cognitive function has marginally improved over time. The r coefficients for AD, MCI, and NL are -0.077, -0.036, and -0.350 respectively. However, by comparing this to Fig. 5a, the average ECogSP score has increased over time, which is completely opposite to the trends of the average ECogPT score. This indicates that the patient reported ECog score is less reliable compared to the study partner reported ECog score, since the very nature of AD is a worsening progression and seldom getting better. This result is consistent with other research findings [7, 29] who found patient reported memory complaints to be less accurate compared to study partner reported complaints. In both Fig. 5a and b, the average of ECog scores for MCI groups is very steady for both study partner and participant, with r coefficients 0.724 and -0.036 respectively.

In Fig. 6a and b, the average ECog scores for participants who at baseline were diagnosed with MCI and had no subsequent change in diagnosis has been plotted side-by-side to MCI participants who have progressed into AD within 48 months. In Fig. 6a, the progression in

Fig. 5 **a** Average ECogPT total score over time for partner. **b** Average ECogPT total score over time for participant

Fig. 6 **a** MCI no change vs MCI to AD, average ECogSP total score. **b** MCI no change vs MCI to AD, average ECogPT total score
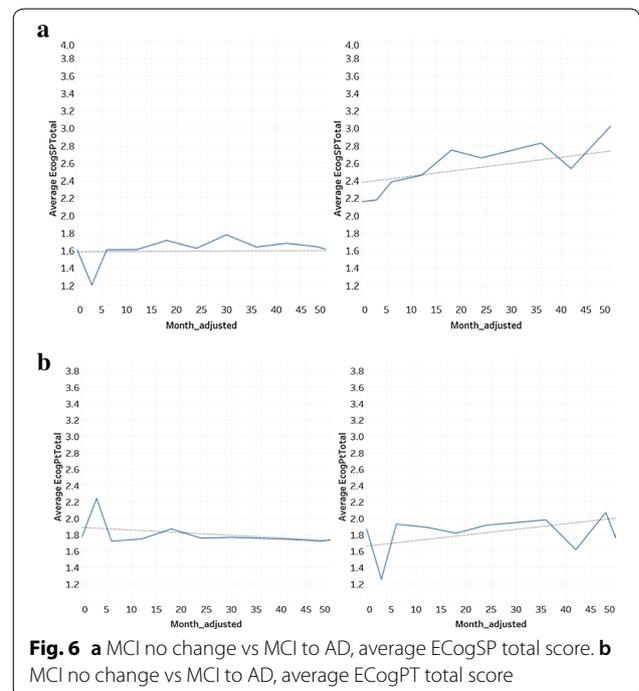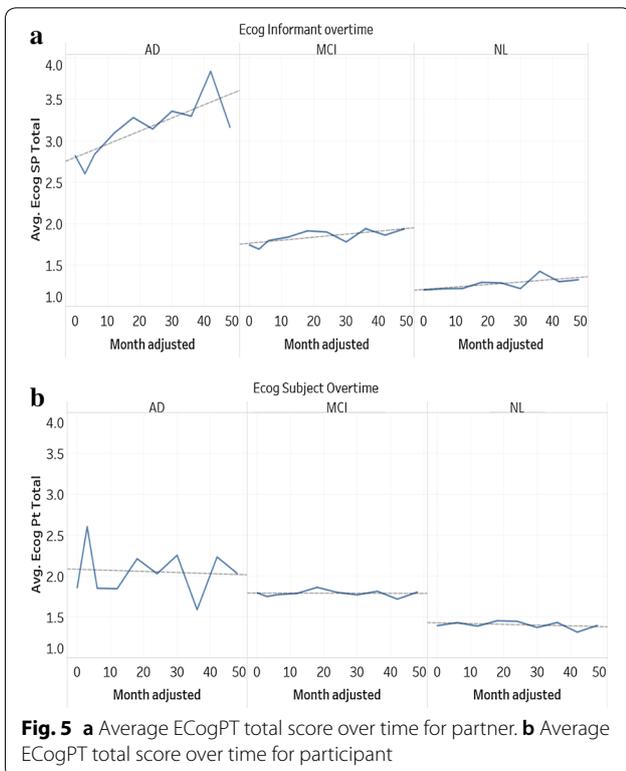
EcogSP score from the "MCI to AD" group can be clearly seen where the average ECogSP score has increased from 2.158 to 2.906 in 48 months with an r coefficient of 0.827, while the MCI "no change" group remains between 1.615 and 1.780 with an r coefficient of 0.487.

However, regarding the average ECogPT score, there are different trends between these two groups. In the MCI "no change" group, the trend is slightly downward and ranging from 1.872–1.722, with an outline score of 2.244 and r coefficient of -0.431. In the "MCI to AD" group, the trend is upward and in the range of 1.879–2.073, with an r coefficient of 0.345 and a few outliers. Those outliers can be explained by a lower visit number in certain months. When comparing Figs. 6a and b, the prior discovery of ECogPT scores tending to be lower than ECogSP scores has stayed true in the MCI group who progressed into AD within 48 months. Yet, the ECogPT score in MCI no change group tends to slightly underestimate their cognitive decline when the ECogSP score remains constant.

It should be noted that there exists a large variation in ECog scores for subjects from all groups, so while taking the average of ECog scores can show the trend for different groups over time, individual participants ECog scores do not always follow these trends.

Machine learning results

To evaluate each version of the ECog test's ability to predict a progression from MCI patients to AD within 48 months, we undertook a machine learning experiment using the Waikato Environment for Knowledge Analysis (WEKA) [18] tool version 3.8.3 on a device with an 2.80 GHz, i7 processer, and 16 GB of RAM.

Three feature sets were created from patients diagnosed with MCI at baseline. Firstly, a feature set consisting of only demographic features: Participant's age, gender, years of education, and race were used as a benchmark. Secondly a feature set with demographic features plus ECogPT Total and ECogPT Mem scores, and thirdly a feature set with demographic features plus ECogSP Total and ECogSP Mem scores. The MCI_Change feature as described in "Dataset description" section was used as the class; since we are limiting to participants diagnosed

with MCI at baseline it has two possible values: "No Change" and "MCI to AD".

RIPPER, PART, C4.5, and Random Forest classification algorithms were used to train a total of 12 predictive models. Default WEKA parameters were used in all cases. The models were evaluated using tenfold cross validation. Using 10 repetitions the mean values of the performance metrics: accuracy, precision, and recall, were recorded for the models derived by the machine learning algorithms.

All classification models trained had unsatisfactory precision and recall rates and couldn't be confidently used as a predictor of MCI to AD progression as shown in Table 1. One of the principal reasons for the low precision and recall rates is the data imbalance issue with respect to the class label—this caused the machine learning algorithms to ignore the minority class, i.e. "MCI to AD". However, the differences in performance for models trained on separate feature sets are interesting and align with the findings of the previously described trend analysis. Models trained on the ECogPT with a demographics feature set were not found to have a significantly different performance than models trained on the benchmark demographics feature set. It should be noted that of the 560 subjects with MCI at baseline, only 132 (24%) progress to an AD diagnosis within 48 months indicating there is an issue of class imbalance. This, along with lack of information in the features, has caused some of the models to classify all instances as "No Change" resulting in biased accuracies of nearly 76% for most of the machine learning algorithms.

To reduce the issue, SMOTE resampling technique was employed on the set of all baseline MCI visits to generate more instances for the minority class; in this case "MCI to AD". This pre-processing operation was completed using WEKA's SMOTE package with all default parameters, aside from the percentage of instances to generate which was set to 200% of the minority class. This resulted in a new dataset of 824 instances of which 396 (48%) belong to the "MCI to AD" class. The machine learning experiment was repeated using the resampled dataset and the results can be seen in Table 2.

Table 1 Performance of classification models

	Demo			Demo + ECogPT			Demo + ECogSP		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
RIPPER	0.7602	0.1078	0.0023	0.7636	0	0.0008	0.7895	0.6071	0.2995
PART	0.7536	0.2458	0.0234	0.7434	0.281	0.0365	0.7695	0.5155	0.2974
C4.5	0.7643	0	0	0.7632	0	0.0008	0.7875	0	0.3091
Random forest	0.7191	0.3451	0.2013	0.7268	0.2667	0.0892	0.7839	0.5882	0.3387

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 9 of 11

Table 2 Performance of classification models after SMOTE

	Demo			Demo + ECogPT			Demo + ECogSP		
	Accuracy	Precision	Recall	Accuracy	Precison	Recall	Accuracy	Precison	Recall
RIPPER	0.6239	0.6137	0.5956	0.6342	0.6257	0.6038	0.6971	0.6929	0.6729
PART	0.6291	0.6002	0.6523	0.6524	0.6213	0.7333	0.7147	0.6544	0.8701
C4	0.6291	0.6046	0.6805	0.653	0.6232	0.7163	0.725	0.683	0.8101
Random forest	0.7432	0.7391	0.7256	0.7339	0.7246	0.7273	0.7834	0.7681	0.7919

After resampling, the precision and recall of models trained on all three feature sets improved greatly. Again, the models trained on the ECogSP with a demographics feature set performed better than the models trained on the ECogPT with demographics and only demographics feature sets. Random Forest when used to process the ECogSP with a demographics feature set produced the highest performing model in terms of accuracy and precision with 0.7834 and 0.7681 respectively, and a recall of 0.7919. The model trained on the same feature set using PART achieved the highest recall of 0.8701 with an accuracy of 0.7147 and recall 0.6544.

Conclusions

AD is one of the leading issues in health care today. While there is currently no cure for AD, once a patient is promptly diagnosed there are treatments that can be followed to improve the patient's living conditions and help them to maintain their independence. It is evident from the correlation analysis presented in this paper that there is a difference between ECogSP and ECogPT scores over time, particularly in participants who are diagnosed with AD at baseline or participants who are diagnosed with MCI at baseline but then progress to AD at a subsequent visit. We found that the average participant reported ECogPT score trends to be slightly down over time in all baseline diagnosis groups: NL, MCI, and AD indicating an improvement in cognitive ability. However, the average study partner reported ECogSP score trends upwards for AD participants and MCI participants who later progress to AD while remaining fairly stable over time for NL participants and MCI participants who do not progress to AD. We also found that there was a significantly different average ECogSP score at baseline between the diagnosis groups compared to a smaller difference for the average ECogPT scores at baseline.

In addition, we have completed a machine learning experiment to evaluate if it is possible to build a model that can predict if a patient with MCI will progress to AD within 48 months based on ECog scores and demographic information. While the models initially created performed poorly in terms of precision and recall, after resampling using SMOTE to reduce class imbalance the results in terms of recall and precision of models increased greatly. Random Forest algorithms derived more competitive predictive models after resampling when compared with the other considered classification algorithms, making it a suitable classifier for predicting progression of AD, at least on the dataset we considered. In all cases, the models that used the ECogSP versions of the test performed significantly better than models trained using the ECogPT version or the demographics alone. This reiterates the finding that the study partner-reported version of the ECog test is a better indicator of progressing AD.

Participants in the ADNI study diagnosed as having MCI or AD tend to report better cognitive ability in the ECog Test than reported about them by a study partner. MCI or AD participants appear to report consistent cognitive ability over time despite the progressive nature of the disease.

By taking the mean values of ECog scores at each time point we lose the individual element of the longitudinal data, therefore a more in-depth approach including longitudinal modelling would be needed to substantiate our findings. This paper could be expanded in future work by using a longitudinal model to evaluate the correlation between ECogPT and ECogSP scores over time. In

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 10 of 11

Table 3 Dataset features

Description	Types	Additional information
RID (roster ID) uniquely identifies every participant	Nominal	1737 participants
Participant ID	Nominal	1737 participants
Visit code, time point when the visit takes place	Numeric	Mean 7.3 visits per participant; (bl is baseline or month 0, m06 is month 6, etc.)
Site ID	Numeric	62 Sites
Participant's age	Numeric	Mean 73.7
Participant's gender	Nominal	Male 957/female 780
Participant's years of education	Numeric	Mean 15.9
Race, Latino vs non-Latino	Nominal	Latino 58; non-Latino 1669; unknown 11
Race, white, black, etc.	Nominal	White 1605; black 77; Asian 29; mixed race 18; Hawaiian 2; Indian 3; unknown 3
Marital Status	Nominal	Married 1311; widowed 206; divorced 150; single 64; unknown 7
Months passed since baseline visit	Numeric	0,3,6…126
Cognitive test—participant everyday cognitive—everyday memory ()	Numeric	Mean 2.07
Cognitive Test—Participant Everyday Cognitive—Total	Numeric	Mean 1.68
Cognitive test—study partner everyday cognitive—everyday memory	Numeric	Mean 2.14
Cognitive test—study partner everyday cognitive—total	Numeric	Mean 1.82
Cognitive test—participant everyday cognitive—everyday memory (baseline)	Numeric	Baseline ECogPT reading, with several missing values
Cognitive test—participant everyday cognitive—total (baseline)	Numeric	Baseline ECogPT reading, with several missing values
Cognitive test—study partner everyday cognitive—everyday memory (baseline)	Numeric	Baseline ECogSP reading, with several missing values
Cognitive test—study partner everyday cognitive—total (baseline)	Numeric	Baseline ECogSP reading, with several missing values
Baseline clinical diagnosis	Nominal	AD 342,EMCI 310, LMCI 562, SMC 106, CN 417,
Clinical diagnosis	Nominal	AD 336, AD to MCI 1, MCI 864, MCI to AD 5, MCI to NL 2, NL 521, NL to MCI 1

addition, the six ECog sub scores which measure individual cognitive domains could also be assessed for both versions of the test.

Appendix
See Table 3.

References
1. ADNI. *Alzheimer's disease neuroimaging initiative*. 2017. https://adni.loni.usc.edu/about/#core-container. Accessed 22 Oct 2019.
2. Bakkour A, Morris JC, Wolk DA, Dickerson BC. The cortical signature of prodromal AD: regional thinning predicts mild AD dementia. Neurology. 2009;72:1048–55.
3. Balsis S, Benge JF, Lowe DA, Geraci L, Doody RS. How do scores on the ADAS-Cog, MMSE, and CDR-SOB correspond? Clin Neuropsychol. 2015;29(7):1002–9.
4. Bates DM, Pinheiro JC. Linear and nonlinear mixed-effects models. Appl Stat Agric. 1998. https://doi.org/10.4148/2475-7772.1273.
5. Bergeron D, Flynn K, Verret L, Poulin S, Bouchard RW, Bocti C, Fülöp T, Lacombe G, Gauthier S, Nasreddine Z, Laforce RJ. Multicenter validation of an MMSE-Mo CA conversion table. J Am Geriatr Soc. 2017;65(5):1067–72.
6. Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.
7. Carr DB, Gray S, Baty J, Morris JC. The value of study partner versus individual's complaints of memory impairment in early dementia. Neurology. 2000;55(11):1724–7.
8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
9. Cohen W. Fast effective rule induction. In: Prieditis A, Russell S, editors. Proceedings of the 12th international conference on machine learning, ICML. Tahoe City: Morgan Kaufmann; 1995. p. 115–23.
10. Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiol Aging. 2011;32(12):2322–e19.
11. Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: The SPARE-AD index. Brain. 2009;132(8):2026–35.
12. Desikan RS, Cabral HJ, Settecase F, Hess CP, Dillon WP, Glastonbury CM, Weiner MW, Schmansky NJ, Salat DH, Fischl B, The Alzheimer's Disease Neuroimaging Initiative. Automated MRI measures predict progression to Alzheimer's disease. Neurobiol Aging. 2010;31(8):1364–74. https://doi.org/10.1016/j.neurobiolaging.2010.04.023.
13. Evans MC, Barnes J, Nielsen C, Kim LG, Clegg SL, Blair M, Leung KK, Douiri A, Boyes RG, Ourselin S, Fox NC. Volume changes in Alzheimer's disease and mild cognitive impairment: cognitive associations. Eur Radiol. 2010;20(3):674–82.

Thabtah *et al. Health Inf Sci Syst*   (2020) 8:24

Page 11 of 11

14. Farias ST, Mungas D, Reed BR, Cahn-Weiner D, Jagust W, Baynes K, DeCarli C. The measurement of everyday cognition (ECog): scale development and psychometric properties. Neuropsychology. 2008;22(4):531.

15. Folstein M, Folstein SE, McHugh P. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189–98. https://doi.org/10.1016/0022-3956(75)90026-6.

16. Frank E, Witten I. Generating accurate rule sets without global optimisation. In: Proceedings of the fifteenth international conference on machine learning, Madison, WI; 1998. p. 144–51.

17. Geuze E, Vermetten E, Bremner JD. MR-based in vivo hippocampal volumetrics: 2. Findings in neuropsychiatric disorders. Mol Psychiatry. 2005;10(2):160–84. https://doi.org/10.1038/sj.mp.4001579.

18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explor Newsl. 2009;11(1):10–8.

19. Ito K, Hutmacher MM, Corrigan BW. Modeling of functional assessment questionnaire (FAQ) as continuous bounded data from the ADNI database. J Pharmacokinet Pharmacodyn. 2012;39(6):601–18.

20. Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 2010;9(1):119–28. https://doi.org/10.1016/S1474-4422(09)70299-6.

21. Jack CR, Lowe VJ, Weigand SD, Wiste HJ, Senjem ML, Knopman DS, Shiung MM, Gunter JL, Boeve BF, Kemp BJ, Weiner M, Petersen RC, The Alzheimer's Disease Neuroimaging Initiative. Serial PIB and MRI in normal, mild cognitive impairment and Alzheimer's disease: implications for sequence of pathological events in Alzheimer's disease. Brain. 2009;132(5):1355–65. https://doi.org/10.1093/brain/awp062.

22. Killiany RJ, Hyman BT, Gomez-Isla T, Moss MB, Kikinis R, Jolesz F, Tanzi R, Jones K, Albert MS. MRI measures of entorhinal cortex vs hippocampus in preclinical AD. Neurology. 2002;58(8):1188–96. https://doi.org/10.1212/WNL.58.8.1188.

23. Ministry of Health NZ. Dementia, treatment. 2018. Retrieved from https://www.health.govt.nz/your-health/conditions-and-treatments/diseases-and-illnesses/dementia. Accessed 4 Jan 2020.

24. Moradi E, Hallikainen I, Hänninen T, Tohka J, Alzheimer's Disease Neuroimaging Initiative. Rey's auditory verbal learning test scores can be predicted from whole brain MRI in Alzheimer's disease. NeuroImage Clinical. 2017;13:415–27.

25. Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am. 2005;15(4):869–xii. https://doi.org/10.1016/j.nic.2005.09.008.

26. Pfeffer RI, Kurosaki TT, Harrah HC, Chance JM, Filos S. Measurement of functional activities in older adults in the community. J Gerontol. 1982;37(3):323–9. https://doi.org/10.1093/geronj/37.3.323.

27. Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc. 1984;79(387):575–83.

28. Quinlan JR. C4. 5: programs for machine learning. Amsterdam: Elsevier; 2014.

29. Rabin LA, Wang C, Katz MJ, Derby CA, Buschke H, Lipton RB. Predicting Alzheimer's Disease: Neuropsychological tests, self-reports, and study partner reports of cognitive difficulties. J Am Geriatr Soc. 2012;60(6):1128–34.

30. Rey A. L'examen psychologique dans les cas d'encéphalopathie traumatique. Arch Psychol. 1941;28:286–340.

31. Rosen W, Mohs R, Davis K. A new rating scale for Alzheimer's disease. Am J Psychiatry. 1984;141(11):1356–64. https://doi.org/10.1176/ajp.141.11.1356.

32. Schmidt M. Rey Auditory verbal learning test: a handbook. Los Angeles, CA: Western Psychological Services; 1996.

33. Schuff N, Woerner N, Boreta L, Kornfield T, Shaw LM, Trojanowski JQ, Thompson PM, Jack CR Jr, The Alzheimer's Disease Neuroimaging Initiative. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. Brain. 2009;132(4):1067–77. https://doi.org/10.1093/brain/awp007.

34. The Alzheimer's Disease Prediction of Longitudinal Evolution (TADPOLE). *TADPOLE-Home*. 2019. https://tadpole.grand-challenge.org/. Accessed 5 Dec 2019.

35. Varon D, Barker W, Loewenstein D, Greig M, Bohorquez A, Santos I, Shen Q, Harper M, Vallejo-Luces D, R. Visual rating and volumetric measurement of medial temporal atrophy in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort: baseline diagnosis and the prediction of MCI outcome. Int J Geriatr Psychiatry. 2015;30(2):192–200. https://doi.org/10.1002/gps.4126.

36. Vemuri P, Wiste H, Weigand S, Knopman D, Trojanowski J, Shaw L, Bernstein MA, Aisen PS, Weiner M, Petersen RC, Jack CR Jr, Alzheimer's Disease Neuroimaging Initiative. Serial MRI and CSF biomarkers in normal aging, MCI, and AD. Neurology. 2010;75(2):143–51. https://doi.org/10.1212/WNL.0b013e3181e7ca82.

37. Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, Harvey D, Jack CR, Jagust W, Liu E, Morris JC, Petersen RC, Saykin AJ, Schmidt ME, Shaw L, Siuciak JA, Soares H, Toga AW, Trojanowski JQ, Si JA. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. Alzheimer's Dement. 2012;8(10):S1–S68. https://doi.org/10.1016/j.nic.2005.09.008.

38. World Health Organization. (2019). *Dementia*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/dementia. Accessed 11 Oct 2019.

39. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–20.