NeuroImage: Clinical 25 (2020) 102121

Contents lists available at ScienceDirect

NeuroImage: Clinical

journal homepage: www.elsevier.com/locate/ynicl

MRI and cognitive scores complement each other to accurately predict Alzheimer's dementia 2 to 7 years before clinical onset

Azar Zandifar[a,b,*], Vladimir S. Fonov[a], Simon Ducharme[a,c], Sylvie Belleville[d,e], D. Louis Collins[a,b], for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University Street, Room WB320, Montreal, QC H3A 2B4, Canada
[b] Department of Biomedical Engineering, McGill University, Montreal, Canada
[c] Department of Psychiatry, McGill University Health Centre, McGill University, Montreal, Canada
[d] Institut Universitaire de Gériatrie de Montréal, Montreal, Canada
[e] Department of Psychology, Centre de Recherche en Neuropsychologie et Cognition, Université de Montréal, Montreal, Canada

A B S T R A C T

Background: Predicting cognitive decline and the eventual onset of dementia in patients with Mild Cognitive Impairment (MCI) is of high value for patient management and potential cohort enrichment in pharmaceutical trials. We used cognitive scores and MRI biomarkers from a single baseline visit to predict the onset of dementia due to AD in an amnestic MCI (aMCI) population over a nine-year follow-up period.
Method: All aMCI subjects from ADNI1, ADNI2, and ADNI-GO with available baseline neurocognitive scores and T1w MRI were included in the study (n = 756). We built a Naïve Bayes classifier for every year over a 9-year follow-up period and tested each one with Leave one out cross validation.
Results: We reached 87% prediction accuracy at five years follow-up with an AUC > 0.85 from two to seven years (peaking at 0.92 at five years). Both neurocognitive scores and MRI biomarkers were needed to make the prognostic models highly sensitive and specific, especially for longer follow-ups. MRI features are more sensitive, while cognitive features bring specificity to the prediction.
Conclusion: Combining cognitive scores and MRI biomarkers yield accurate prediction years before onset of dementia. Such a tool may be helpful in selecting patients that would most benefit from lifestyle changes, and eventually early treatments that would slow cognitive decline and delay the onset of dementia.

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disease (Association, 2013; Jack et al., 2011; McKhann et al., 2011). The prodromal stage of AD dementia, known as Mild Cognitive Impairment (MCI), is characterized by the gradual onset and evolution of cognitive impairment beyond the levels expected for age and education of the individual, but without interfering with a patient's everyday life (Petersen et al., 1999). MCI patients with memory problems as their main symptom are known as "amnestic MCI", from which 10% − 15% are reported to progress to clinically probable AD each year (Petersen, 2009). Since not all amnestic MCI (aMCI) patients progress to AD, predicting if and when a subject with aMCI will have future dementia will enable enrichment for clinical trials. More importantly, Kivipelto's group (Ngandu et al., 2015) has demonstrated the potential benefit of combining diet, exercise and cognitive training to prevent cognitive decline. Early detection of prodromal disease (e.g., 5–10y in advance) is key to intervene before the onset of cognitive decline due to irreversible neurodegeneration.

Jack et al. (2010) well-known hypothetical biomarker model and its successors (Sperling et al., 2011) describe the dynamics of biomarkers during AD pathological process. The model proposes that cognition and structural brain atrophy change with the sharpest slope in the MCI stage (Sperling et al., 2011), making them potentially the most sensitive early biomarkers of progression from aMCI to AD. According to this model, anatomical atrophy, which can be captured by volumetric MRI, begins ahead of cognitive decline (Sperling et al., 2011). Specifically, the entorhinal and hippocampal areas are known to be among the first affected (Jack et al., 1997; Braak and Braak, 1991). Factoring the sharp slope of atrophy, early stage volume loss and its widespread

* Corresponding author at: McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University Street, Room WB320, Montreal, QC H3A 2B4, Canada.
E-mail addresses: azar.zandifar@mail.mcgill.ca (A. Zandifar), vladimir.fonov@mcgill.ca (V.S. Fonov), simon.ducharme@mcgill.ca (S. Ducharme), sylvie.belleville@umontreal.ca (S. Belleville), louis.collins@mcgill.ca (D.L. Collins).

[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-ontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

https://doi.org/10.1016/j.nicl.2019.102121
Received 17 March 2019; Received in revised form 17 November 2019; Accepted 10 December 2019
Available online 16 December 2019

availability, MRI morphometric analyses is a promising candidate biomarker for earlier prediction.

To measure AD-related morphological changes in brain anatomy, we have developed the Scores by Nonlocal Image Patch Estimator (SNIPE) for both hippocampal and entorhinal areas (Coupé et al., 2012b). SNIPE is a similarity metric that measures structural similarity to either a library of cognitively normal subjects or a library of patients with Alzheimer's dementia. We previously showed that using only SNIPE scores for hippocampus, plus age and sex, one can reach an overall accuracy of 71% for prediction of progression from aMCI to dementia over a 3y follow-up period (Coupé et al., 2012b). In a more recent work looking at a cohort of cognitively healthy older individuals, we showed that our MRI-driven SNIPE biomarker was sensitive to AD-related changes in a cognitively normal cohort on average seven years before clinical diagnosis of AD dementia (Coupe et al., 2015).

While volumetric MRI measures have good prediction value, we hypothesize that prediction performance can be improved using complementary information of other features, in particular performance on standardized cognitive tests. Indeed, previous studies showed that the patient's current cognitive state can also predict future cognitive decline in aMCI (Belleville, 2017, Belleville et al., 2008). The ACE-R (Addenbrooke's Cognitive Examination - Revised) (Lischka et al., 2012), MOCA (Montreal Cognitive Assessment) (Julayanont et al., 2014), verbal memory measures and many language tests (Belleville, 2017) all have shown promising performance to predict future dementia in persons with aMCI. However, some models may have difficulty with short term prediction as they found the likelihood of false negatives was increased, resulting in decreased sensitivity to imminent onset of AD (Belleville, 2017). Furthermore, these studies either suffer from relatively short follow-up periods, or have not used a combination of different scores and biomarkers to benefit from their complementary information. While a more recent study shows that combining both cortical thinning measures and cognitive scores increases prediction accuracy (Peters et al., 2014), this study too had a relatively short follow-up period, and a limited number of subjects and did not investigate the complementary effect of the different features; namely how and to what extent each feature set contributes in the performance of the predictive model (Peters et al., 2014).

While both SNIPE and cognition have shown potential to predict conversion, it is uncertain if a combination of both measures would lead to superior accuracy. In this study we have two main goals. First, we investigate the ability of our model to predict progression to AD based only on the baseline MRI and cognitive information over follow-up periods ranging from one to nine years. Second, we investigate the change over time in predictive power of each feature to investigate whether, as previously suggested in Jack's model (Jack et al., 2010), the MRI-driven biomarkers are better predictors than neurocognitive scores for longer follow-up periods (as atrophy is hypothesized to precede cognitive decline), and whether they can capture AD-related abnormality before cognitive scores do. This study further investigates the effect of combining neurocognitive scores and MRI makers throughout different follow-up periods using a large dataset.

2. Methods

In this study, we train a Naïve Bayes classifier to predict the future diagnosis of dementia in patients with aMCI in the ADNI1, ADNI2 and ADNI-GO datasets. Our feature set contains age, sex, years of education, neurocognitive (Alzheimer's Disease Assessment Scale – Cognitive subscale [ADAS-cog], Rey Auditory Verbal Learning Task [RAVLT], Mini-Mental State Examination [MMSE]), MRI-based scores (SNIPE scores for hippocampus and entorhinal cortex), and disease severity (CDR-SB [Clinical Dementia Rating Sum of Boxes]) from baseline data that are used as input to the classifier. The classifier then attempts to predict the future diagnosis for each patient. To train the classifier, we need patient diagnostic information collected at later time points during the study. The follow-up period is the time interval from baseline to later time points for which we know the state of the subject: i.e., whether that subject has progressed to dementia or has maintained the aMCI stage.

2.1. Dataset

2.1.1. Dataset selection

Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of aMCI and early Alzheimer's disease (AD). The study was approved by each participant sites' Review Board. A written informed consent was obtained from each participant at the time of enrolment that included permission for analysis and data sharing.

We used all amnestic MCI subjects from ADNI1, ADNI2, and ADNI-GO for which the baseline cognitive scores and T1 MRI were present (n = 756). We used MRI and cognitive data from the baseline visit to predict the future clinical diagnostic status at follow-up at 12, 24, 36, 48, 60, 72, 84, and 108 months (See Table 1). (While 120-month data is available, too small a number of subjects are available for model testing and validation.)

At each visit of the ADNI study, patients are evaluated for AD based on NINCDS-ADRDA criteria (McKhann et al., 2011). Amnestic MCI participants are identified as those that have reported a subjective memory concern either autonomously or via an informant or clinician; have abnormal performance on the Wechsler Memory Scale Logical Memory II test, however activities of daily living are preserved (CDR = 0.5) and they do not meet criteria for a dementia diagnosis.

2.1.2. Dataset labeling

The subjects were labeled either stable MCI (sMCI) or progressive MCI (pMCI) label based on the difference in diagnostic status between the baseline visit and each follow-up time point. For example, at 24 months, we included all aMCI subjects for which the diagnostic state at

Table 1
Dataset information based on follow-up duration.

	Baseline	12 months	24 months	36 months	48 months	60 months	72 months	84 months	96 months	108 months
pMCI subjects (#)	756 total	101	174	174	127	70	53	32	23	17
sMCI subjects (#)		619	431	325	200	97	50	44	33	21
pMCI:sMCI ratio		0.163	0.404	0.535	0.635	0.722	1.060	0.727	0.697	0.810
Mean Age at baseline (Standard Deviation)	73 (7)	73 (7)	73 (7)	73 (7)	72 (7)	72 (7)	74 (7)	74 (7)	74 (7)	73 (7)
Female Percentage (%)	40	39	40	38	38	35	35	34	29	34
Median education (First, third quartile)	16 (14,18)	16 (14,18)	16 (14,18)	16 (14,18)	16 (14,18)	16 (13,18)	16 (13,18)	16 (12,18)	16 (12,18)	16 (13,18)
Median MMSE (First, third quartile)	28 (26,29)	28 (26,29)	28 (26,29)	28 (26,29)	28 (27,29)	28 (27,29)	27 (26,29)	28 (27,29)	28 (27,29)	28 (27,29)

Note: sMCI shows stable MCI subjects, while progressive MCI population are referred to as pMCI. All information reported is based on baseline visit for each follow-up cohort.

24 months was present, and their clinical status remained unchanged or worsen during the follow-up period. The subjects who maintained their baseline amnestic MCI state were labeled sMCI. Subjects that progressed to a dementia due to AD diagnosis at any time up to and including the follow-up time point received the pMCI label. We do not consider later time points, as our goal is to match the clinical status at the given time point. It is important to note that some subjects may return to a normal cognitive state at later follow-up visits. However, we do not use this information at the earlier time point, as this information is not available at that time and we wish to evaluate the classifier, as it would perform in a prospective context. When a subject did not have a diagnostic label for a time point, that subject was dropped from the analysis for that follow-up time point only. The detailed aMCI dataset information is given in Table 1. During long, large studies such as ADNI drop-outs can occur due to many reasons, which could potentially bias the outcome of studies. In order to prevent such biases, we limited ourselves to subjects for whom the clinical status is known and data is present in both clinical groups (stable and progressive MCI at every time point). We believe that such a strategy is more robust to dropout biases in comparison to including the most recent clinical status as the clinical status for each given follow-up period. Therefore, the strategy used here to categorize the data prevents the bias that could have been introduced to the analyses due to the clinical status of the subjects who dropped out from the study.

2.2. Measures

2.2.1. MRI derived biomarker

Hippocampal and entorhinal SNIPE grading scores were used as the only MRI biomarker features in the predictive classifier. The SNIPE score is described fully in (Coupe et al., 2015, 2012a). In short, after preprocessing with our in-house pipeline that includes denoising (Coupe et al., 2008), N3 inhomogeneity correction (Sled et al., 1998), linear intensity normalization based on histogram matching between the image and the average template, and affine registration to ICBM152 template space with $1 \times 1 \times 1$ mm^3 resolution (Collins et al., 1994), SNIPE assigns a similarity metric to each voxel, which shows how much that voxel's neighbourhood resembles the anatomy of either, a group of patients with Alzheimer's dementia or a group of normal controls. The final SNIPE score is an average of all the voxels in the desired anatomical structure in each hemisphere (Coupé et al., 2015). For this study the SNIPE scores are corrected for age and sex using the method presented in Dukart et al. (2011). The scores are corrected based on a linear regression model fitted only on the cognitively normal population to correct for the effect of normal aging and preserve the effect of disease-related changes. All processed MRI was submitted to visual quality control. No datasets needed to be excluded for insufficient quality.

2.2.2. Neurocognitive scores

As mentioned before, previous studies showed that the prognostic value of neurocognitive scores can vary depending on remaining time to future onset of dementia (Belleville, 2017). Here, we included the baseline neurocognitive scores available within the ADNI study. These included the total score of ADAS-cog (both ADAS-cog-11 and ADAS-cog-13), the Rey Auditory Verbal Learning Task (RAVLT) scores of immediate recall, learning and forgetting, and MMSE. RAVLT is administered by presenting a list of 15 words across five consecutive trials. The performance of the participant is measured by the ability to recall the words immediately after each trial (immediate – recall), or recalling the words after presenting the participant with different set of words and a time delay (learning), and number of the words that the participant forgets from the initial list (forgetting). We also included baseline CDR-SB. The values are corrected for age, sex and education using the method presented by Dukart et al. (2011).

2.3. Procedures and statistical analysis

2.3.1. Classification and validation

We trained three classifiers based on the following features: first, using only four MRI-driven biomarkers (both left and right entorhinal and hippocampal SNIPE score) and age, sex and years of education; second, using only seven neurocognitive scores plus age, sex and education; and third, using all MRI driven and cognitive scores together plus age, sex and education. All features are drawn only from the baseline visit. For each follow-up, we trained a specific Naïve Bayes classifier, for which the baseline information is used as features and the stable/progressing state of a patient at that specific follow-up period is regarded as the desired output. All classifiers are tested with a Leave One Out (LOO) cross-validation technique. This means that in each step, the classifier is trained on all the subjects from the dataset for that specific follow-up except one specific subject that is used as the test subject. This process is repeated so that all subjects are used for testing. Therefore, we have a completely separate testing and training sets.

2.3.2. Classification accuracy, sensitivity, specificity and area under the receiver operating curve (AUC)

The classification performance at each time point is measured by classification accuracy, sensitivity and specificity for each follow-up period. To make the comparison between the metrics for different settings (using only MRI-driven scores, only using neurocognitive scores, or using both sets of features) feasible and to gain a more robust sense of the classification performance, we measured the classification accuracy, sensitivity and specificity by sampling 85 percent of data without replacement for each iteration and we repeated the performance calculation procedure for 200 iterations. That is, for each iteration, we ran a LOO procedure only using 85% of the population. The final accuracy, sensitivity and specificity shown are the mean of the calculated metrics and the standard deviations is shown with errorbars.

As a measure that is robust to unbalanced classes, we also report the AUC for the classifier trained with all the features to better show our model performance.

2.3.3. Feature importance

To determine the relative importance of the different features over the 9-year follow-up period, we use a metric similar to effect size, customized for Naïve Bayes classification. In summary, we compare the different features based on their importance for our classifier at each point over time in the follow-up period (see the supplementary material for details on how feature importance is computed). This shows how the features gain or lose importance over time and is related to the disease progression.

3. Results

3.1. Classification performance

The accuracy, sensitivity and specificity at each follow-up time point are plotted in Fig. 1 for the three classifier scenarios: using only MRI features, using only neurocognitive scores and using both MRI and neurocognitive features together as inputs. From 24 months onwards, the classification accuracy using both sets of features are better than using only MRI features or only neurocognitive features. MRI features are more sensitive (Fig. 1, middle), while neurocognitive features bring specificity to the prediction (Fig. 1, right).

Statistical comparisons were made between accuracies, sensitivities and specificities of all the classifiers with different feature sets. Models were compared when trained using combined feature sets (i.e., MRI and neurocognitive scores) and with each one of them alone. An ANOVA with post-hoc pairwise Tukey test with 95% confidence level shows that the combined model significantly performs better than the models

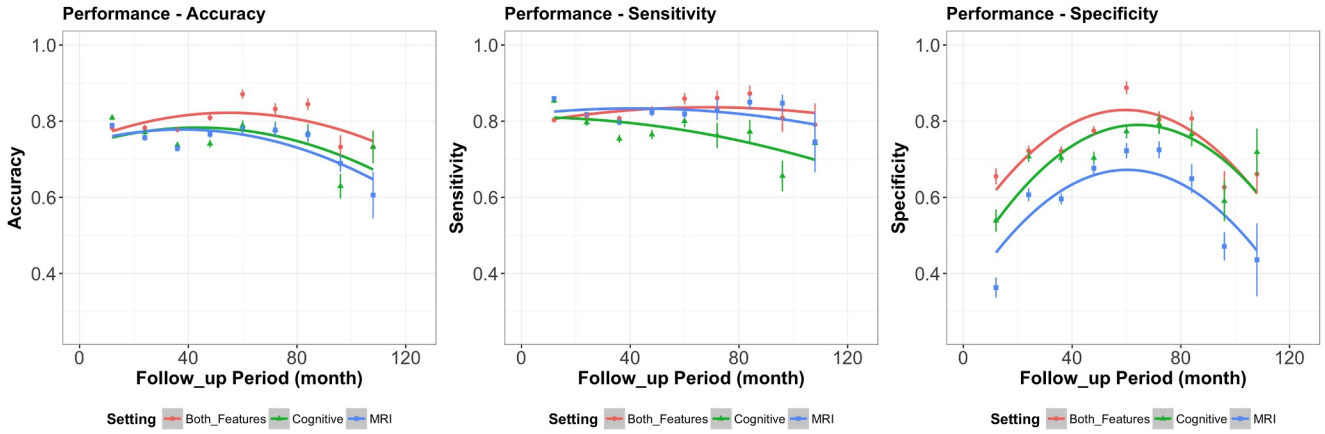

Fig. 1. Accuracy, sensitivity and specificity for classifier performance considering from 12 to 108 months of follow-up.
Data points show the mean metric calculated in each follow-up period from 200 bootstrap samples of 85% of the data. Error bars show the associated standard deviation. The curves show the best second-degree fit.

trained on each feature set alone in all follow-up periods ($p < 2 \times 10^{-5}$) except for the accuracies of the combined model and the model trained with only neurocognitive scores at 108 months.

As stated in the methods section, data imbalance may cause bias towards one of the classes, thus affecting the classification results. For example, at 12mo, there are only 101 pMCI and 619 sMCI. A simple classifier that assigns all the test cases to the majority class would obtain 86% accuracy by assigning all subjects to sMCI class. We therefore measured AUC, which is a more robust metric against data imbalance. The AUC curve in Fig. 2 shows the best second-degree fit (Akaike Information Criterion yielded -34.31 for second order vs. -19.91 for a linear model). The AUC plot over time shows an inverted-U pattern where the maximum performance happens around 60 months ($AUC = 0.92$) with accuracy, sensitivity and specificity of 87%, 86%, and 89%, respectively. The model is robust with AUC values >0.85 for follow-up periods from 24 to 84 months.

The inverted-U pattern in the AUC plot (Fig. 2) is counterintuitive as baseline visits should be more informative for early follow-up periods and less informative for longer follow-up periods and, therefore, model performance should fall monotonically in time. Indeed, the sensitivity profile drops over time for the classification models using only MRI or only cognition features but remains relatively flat for the combined model. The inverted-U pattern of the AUC curve (Fig. 2) is driven by the inverted-U shape of the specificity plot (Fig. 1), indicating that the classifier does not properly identify the negatives (sMCI) at the earlier follow-up time points.

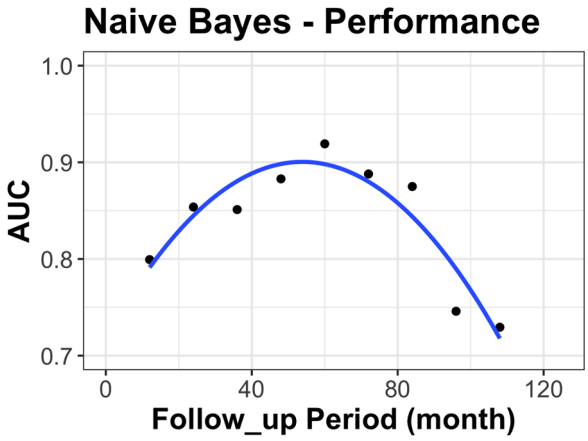

Fig. 2. Area under the receiver operating curve (AUC) for follow-ups from 12 to 108 months.
Data points show the AUC metric value at each follow-up period.

Early follow-ups are prone to more false positives. We hypothesize that this happens due to the fact that shorter follow-ups don't allow enough time for all prodromal AD subjects to progress to a clinical dementia diagnosis. This means that these false positives are individuals who do not change categorical diagnostic categories in shorter follow-ups even though the disease is progressing and will reach the dementia stage at a later time point. To test this hypothesis, one could show that the conversion rate in this group is greater than the average MCI population. From the 109 false positives at 12 months, 70 (64%) converted to dementia, with more than 43% converting within one year (47 converted within one year, 11 within two years, 8 within three years and 4 in four years or more). This is much higher than the expected yearly rate of progression of 10%–15% (Petersen, 2009), indicating that many of these false-positives are in fact true progressors at a later follow-up time.

3.2. Feature importance

Importance of each feature in predicting disease progression at each follow-up is shown in Fig. 3. For illustration, we averaged importance over ADAS-cog-11 and ADAS-cog-13, RAVLT learning, forgetting and immediate, hippocampal SNIPE score for left and right, entorhinal SNIPE score for left and right, and made one score out of each set. This graph shows the value of each feature in making the final decision when the classifier has access to both neurocognitive and MRI data. ADAS-cog shows the greatest importance over the entire follow-up period, followed by hippocampal SNIPE scores.

While all features lose AD-related sensitivity for very long follow-ups, the peak importance for each one happens at a different follow-up period. The later the peak, the more sensitive a feature would be to earlier detection of the disease. Meaning, the farther a peak happens relative to the baseline, the more sensitive is the corresponding feature to the very early patterns of AD-related pathology. If used alone, entorhinal cortex shows the earliest possible detection performance at 68 months and is among the first biomarkers to show abnormality. While the hippocampal grading score has a peak at 39 months, it shows a higher importance for all the follow-up periods from one to seven years compared to the entorhinal cortex SNIPE score. Similarly, by peaking at 62 months, the RAVLT neurocognitive scores are more sensitive to early AD changes in comparison to the more general screening test MMSE and the disease severity marker CDR-SB (performance peak at 47 and 19 months, respectively). Having multiple biomarkers available enables each classifier to select the most appropriate data to achieve high sensitivity and specificity for all follow-up periods.

The demographic information shows very low importance in

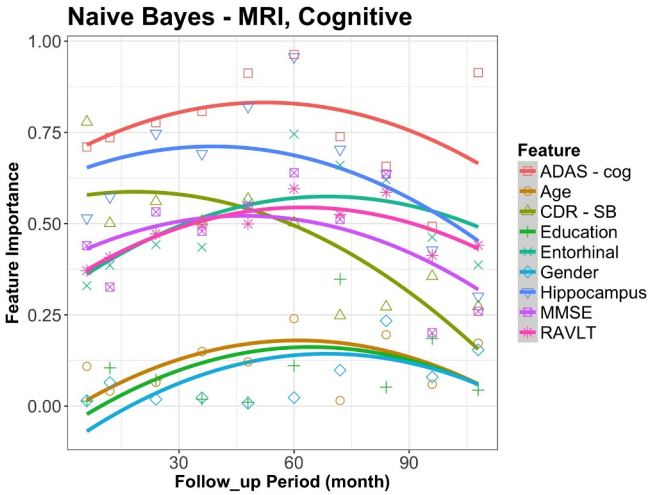

Fig. 3. Feature importance for each follow-up period.
Features show both neurocognitive and MRI-driven scores. The curves show the best second-degree fit given the metric for the 9-year follow-ups.

comparison to all the other features (Fig. 3). In fact, their effect is negligible with respect to the other markers. However, they show higher importance for late follow-ups rather than short ones (i.e., peaks at 61, 68 and 64 months for age, sex and education, respectively).

4. Discussion

In this study, we evaluated a prognostic model to predict the onset of dementia using baseline neurocognitive and MRI-driven features from ADNI subjects as a function of follow-up period.

This study showed that both MRI and neurocognitive information are needed to make the most accurate decision on future progressors. Importantly, they contribute differently to overall accuracy. This observation has been shown previously in (Peters et al., 2014) with a smaller cohort and a single, shorter follow-up period. In the present study, we investigated how the different categories of features contribute to the final performance of the model. Furthermore, this study investigated the effect of different follow-up periods on performance of such a model. The MRI-driven features are more sensitive than neurocognitive test scores, but neurocognitive scores are more specific than MRI features. In other words, the MRI derived biomarkers are highly sensitive to the morphological pattern of abnormal brain aging (i.e., to identify at-risk subjects), while preservation of cognition function contributes to identify persons who will not progress over time despite MRI changes. This would explain why using both make an optimally accurate, sensitive, and specific prognostic tool.

One of the observations in this study was that the AUC follows an inverted-U shape over the 9-year follow-up period: at the beginning, AUC rises with time, reaches a maximum, and then falls again. It has been previously observed that a model trained based on neurocognitive scores can show low predictive ability for shorter follow-up periods (Belleville, 2017). This study reported low sensitivity in shorter follow-ups using some neurocognitive scores such as semantic fluency tasks (Belleville, 2017). By combining many of the neurocognitive scores available in ADNI with MRI features we maintained high sensitivity throughout the 9-year follow-up period. The study postulates that in shorter follow-ups, performance is adversely affected by late converters (i.e., individuals whose progression is not sufficient at the one-year time point to change diagnostic category but reach the dementia stage later). However, to the best of our knowledge, no previous study has shown the late-converters hypothesis and their effect on model performance using a data-driven approach. Our investigation with short follow-ups showed that the majority of false positives are in fact late convertors. Indeed, 42% of the false positives at 12 months convert to dementia within the following year. This is almost three times the rate of conversion to dementia found in the average MCI population (Petersen, 2009). More specifically, when our model falsely predicts the conversion to dementia at 12 months, it is likely that there is pathophysiological progression in the brain, but that more time is needed for the disease to reach the dementia stage. Another possibility might be related to cognitive reserve (Katzman et al., 1988, 1993; Stern et al., 1994). The cognitive reserve hypothesis suggests that at a particular level of AD pathology, different individuals manifest different levels of clinical symptoms of dementia. It has been shown that highly educated individuals are less likely to manifest clinical symptoms of dementia compared to less-educated individuals (Katzman et al., 1988, 1993; Stern et al., 1994). However, in our study, there was no difference in education levels between the false positive and true positive groups at 12 months, which argues against this explanation. However, we believe that considering the limited variability in education level of ADNI data, the effect of cognitive reserve cannot be effectively investigated using ADNI dataset.

Furthermore, our study showed that features which are highly sensitive for short follow-ups – for instance disease severity metric (CDR-SB) - can lose importance during longer follow-ups, while some other features may show low sensitivity to early stage changes and gain importance in longer follow-ups like the entorhinal cortex SNIPE grading and RAVLT score. Furthermore, certain baseline neurocognitive scores, like RAVLT, are better than others for longer follow-ups. This confirms a previous study that showed that verbal memory measures and language tests have high predictive value for progression to dementia during the MCI period (Belleville, 2017; Braak and Braak, 1995). Namely, a simple bedside test like the MMSE still has reasonable accuracy to predict conversion.

For mid-range time periods, the combined MRI and cognitive biomarker model closely predicts the follow-up clinical diagnosis. Our AUC reaches 0.92 for five-year follow up, which, to the best of our knowledge, is the highest AUC reported for five-year follow-up for any predictive method in AD.

This study is not without limitations. While we used the large publicly available multi-site ADNI dataset, the longest follow-up periods (e.g., 96 and 108 months) would benefit from a larger subject pool as the number of subjects was highly constrained by the availability of diagnosis, study data, and availability of MRI at baseline. Since the dataset size was limited in longer follow-ups, we decided to use LOO cross-validation. While the method keeps the test and training sets separate, a less data conservative variant, such as 10-fold cross-validation, could be used in case of availability of larger datasets. Furthermore, for the same reason, we decided to include MRI scans with both 1.5T and 3T field strengths to our dataset. To account for this variability, we parameterised our pre-processing module to decrease the variability in the dataset. In addition, a study showed that the pattern of hippocampal atrophy in AD could be similarly observed using both 1.5 T and 3T (Chow et al., 2015). However, using unique field strength for the analyses could decrease the dataset size, which was avoided due to small number of subjects specifically in longer follow-ups.

In addition, our dataset includes patients only with amnestic MCI who maintained their cognitive status or progressed to AD dementia; therefore we cannot determine the accuracy of our model for Alzheimer's disease beyond the typical clinical presentation or aMCI subjects who revert to cognitively normal status or progress with a non-AD etiology. A more heterogeneous population, more representative of the clinical population with prospective follow-up would be needed to evaluate these tools before they could be used in the clinic. Finally, we do not compare our results to those that use molecular imaging such as amyloid-PET (Mathotaarachchi et al., 2017). In future work, we will combine amyloid imaging with our MRI and cognitive features.

5. Conclusion

We have demonstrated that combining MRI features with neurocognitive test results from a baseline visit can be used to predict the onset of dementia in a large cohort over periods up to 9 years, but with maximal accuracy up to five years. Our study showed that MRI-driven features and Neurocognitive scores complement each other, resulting in a robust predictive method that is both sensitive and specific. We showed that MRI-driven features are more sensitive predictor of early AD patterns, while cognitive scores bring specificity to the prediction.

The two year FINGER trial (Ngandu et al., 2015) demonstrated significant benefit of combined diet, exercise and cognitive training interventions to prevent cognitive decline in cognitively at risk adults. Early detection (e.g., 5–10y) is key to intervene to prevent irreversible neurodegeneration and the onset of cognitive decline. In addition, the uncertainty of future progression is a major source of anxiety for patients, therefore any tool that increase the accuracy of course prediction would be of great benefit. With additional validation, this tool could be useful in the clinic for better patient management and for cohort enrichment in clinical trials of new treatments for AD.

CRediT authorship contribution statement

Azar Zandifar: Conceptualization, Visualization, Formal analysis, Software, Data curation, Writing - original draft, Writing - review & editing. **Vladimir S. Fonov:** Formal analysis, Software, Writing - review & editing. **Simon Ducharme:** Software, Writing - review & editing. **Sylvie Belleville:** Software, Writing - review & editing. **D. Louis Collins:** Software, Conceptualization, Supervision, Visualization, Writing - review & editing.

Acknowledgment

This work was supported by grants from the Canadian Institutes of Health Research (MOP-111169), les Fonds de Research Santé Quebec Pfizer Innovation fund, and an NSERC CREATE grant (4140438 - 2012). We would like to acknowledge funding from the Famille Louise & André Charron.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California.

Dr. Ducharme reports no conflicts of interest related to this study. Dr. Ducharme receives salary funding from the *Fonds de Recherche du Québec- Santé*.

Dr. Collins reports no conflicts of interest related to this study. Dr. Collins provides training and consulting to NeuroRx and is co-founder of True Positive Medical Devices.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.nicl.2019.102121.

References

Association, A.P., 2013. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Publishing.

Belleville, S., et al., 2017. Neuropsychological measures that predict progression from mild cognitive impairment to Alzheimer's type dementia in older adults: a systematic review and meta-analysis. Neuropsychol. Rev. 328–353.

Belleville, S., Sylvain-Roy, S., de Boysson, C., Menard, M.C., 2008. Characterizing the memory changes in persons with mild cognitive impairment. Prog. Brain Res. 169, 365–375.

Braak, H., Braak, E., 1991. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 82 (4), 239–259.

Braak, H., Braak, E., 1995. Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging 16 (3), 271–278.

Chow, N., Hwang, K.S., Hurtz, S., Green, A.E., Somme, J.H., Thompson, P.M., et al., 2015. Comparing 3T and 1.5 T MRI for mapping hippocampal atrophy in the Alzheimer's disease neuroimaging initiative. Am. J. Neuroradiol. 36 (4), 653–660.

Collins, D.L., et al., 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. J. Comput. Assist. Tomogr. 18 (2), 192–205.

Coupe, P., et al., 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441.

Coupé, P., et al., 2012a. Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. Neuroimage 59 (4), 3736–3747.

Coupé, P., et al., 2012b. Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage Clin. 1 (1), 141–152.

Coupe, P., et al., 2015. Detection of Alzheimer's disease signature in MR images seven years before conversion to dementia: toward an early individual prognosis. Hum. Brain Mapp. 36 (12), 4758–4770.

Dukart, J., et al., 2011. Age correction in dementia–matching to a healthy brain. PLoS One 6 (7), e22193.

Jack, C.R., et al., 1997. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. Neurology 49 (3), 786–794.

Jack, C.R., et al., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol. 9 (1), 119–128.

Jack, C.R., et al., 2011. Introduction to the recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. J. Alzheimer's Assoc. 7 (3), 257–262.

Julayanont, P., et al., 2014. Montreal cognitive assessment memory index score (MoCA-MIS) as a predictor of conversion from mild cognitive impairment to Alzheimer's disease. J. Am. Geriatr. Soc. 62 (4), 679–684.

Katzman, R., 1993. Education and the prevalence of dementia and Alzheimer's disease. Neurology 13–20.

Katzman, R., Terry, R., DeTeresa, R., Brown, T., Davies, P., Fuld, P., et al., 1988. Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. Ann. Neurol. 23 (2), 138–144.

Lischka, A.R., Mendelsohn, M., Overend, T., Forbes, D., 2012. A systematic review of screening tools for predicting the development of dementia. Can. J. Aging 31 (3), 295–311.

Mathotaarachchi, S., et al., 2017. Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiol. Aging 59, 80–90.

McKhann, G.M., et al., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimers Dement. 7 (3), 263–269.

Ngandu, T., Lehtisalo, J., Solomon, A., Levälahti, E., Ahtiluoto, S., Antikainen, R., et al., 2015. A 2 year multidomain intervention of diet, exercise, cognitive training, and vascular risk monitoring versus control to prevent cognitive decline in at-risk elderly people (FINGER): a randomised controlled trial. Lancet 385 (9984), 2255–2263.

Peters, F., Villeneuve, S., Belleville, S., 2014. Predicting progression to dementia in elderly subjects with mild cognitive impairment using both cognitive and neuroimaging predictors. J. Alzheimer's Dis. 38 (2), 307–318.

Petersen, R.C., 2009. Early diagnosis of Alzheimer's disease: is MCI too late? Curr. Alzheimer Res. 6 (4), 324–330.

Petersen, R.C., Smith, G.E., Waring, S.C., Ivnik, R.J., Tangalos, E.G., Kokmen, E., 1999. Mild cognitive impairment: clinical characterization and outcome. Arch. Neurol. 56 (3), 303–308.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Sperling, R.A., Aisen, P.S., Beckett, L.A., Bennett, D.A., Craft, S., Fagan, A.M., et al., 2011. Toward defining the preclinical stages of Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. J. Alzheimer's Assoc. 7 (3), 280–292.

Stern, Y., et al., 1994. Influence of education and occupation on the incidence of Alzheimer's disease. JAMA 271 (13), 1004–1010.