Towards a reliable prediction of conversion from Mild Cognitive Impairment to Alzheimer's Disease: stepwise learning using time windows

Telma Pereira[*]
[*]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
telma.pereira@tecnico.ulisboa.pt

Francisco L. Ferreira[*]
francisco.lourenco.ferreira@tecnico.ulisboa.pt

Manuela Guerreiro[†]
[†]Laboratório de Neurociências, IMM and Faculdade de Medicina, Universidade de Lisboa, Portugal
mmgguerreiro@gmail.com

Alexandre de Mendonça[†]
mendonca@medicina.ulisboa.pt

Sara C. Madeira[+]
[+]LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
sacmadeira@ciencias.ulisboa.pt

for the Alzheimer's Disease Neuroimaging Initiative[1]

ABSTRACT

Predicting progression from a stage of Mild Cognitive Impairment to Alzheimer's disease is a major pursuit in current dementia research. As a result, many prognostic models have emerged with the goal of supporting clinical decisions. Despite the efforts, the clinical application of such models has been hampered by: 1) the lack of a reliable assessment of the uncertainty of each prediction, and 2) not knowing the time to conversion. It is paramount for clinicians to know how much they can rely on the prediction made for a given patient (conversion or no conversion), and the time windows in case of conversion, in order to timely adjust the treatments. We propose a supervised learning approach using Conformal Prediction and a stepwise learning approach, where the learning model first predicts whether a patient converts to dementia, or remains stable, and then predicts the more likely progression window (short-term or long-term conversion). We used data from ADNI to test the approach and predict conversion within time windows of up to 2 years (short-term converter) and 2 to 4 years (long-term converter). The exploratory results are promising but compromised by the small number of examples for the long-term converting patients, available for training.

Keywords

Conformal Predictors, Confidence Estimation, Mild Cognitive Impairment, Alzheimer's Disease, Prognostic Prediction

1. INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease with devastating effect on patients and their families, and a huge socio-economic impact in modern societies. Nowadays, more than 30 million people suffer from AD worldwide and its prevalence is expected to triple by 2050 [18]. It is therefore paramount to understand AD and its progression, not only to guide clinical decisions and manage patients and families' expectations, but also to develop new effective treatments.

Mild Cognitive Impairment (MCI) is considered as a transitive stage between healthy aging and dementia [18], where patients have cognitive complaints not interfering with daily live activities [18]. These patients are more likely to develop AD [11]. In this context, studying progression from MCI to dementia is a major challenge in current medical research [1, 21]. Whilst there is no treatment to revert AD's symptoms, an early prognostic of dementia is nevertheless useful to help clinicians taking decisions about their patients' possible treatment and to timely adjust medical appointments.

By following different approaches and using different types of data (biological markers and/or neuropsychological data), researchers have sought for robust prognostic models, to guide clinical decisions by means of a medical decision support system to be used in clinical settings. This clinical decision support system would predict the most likely prognostic for a new MCI patient based on the past history of a cohort of patients with known diagnostics.

Despite the advances made in prognostic prediction for MCI patients, where machine learning techniques achieved promising results [13], some issues have hampered its practical use in clinical settings:

1) *Lack of trustworthy prognostic prediction.* Most prognostic models only produce bare predictions, without providing any assessment of the uncertainty on each prediction. This is a major disadvantage as for clinicians it is paramount to know

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

how much can they trust the prognostic predicted for a given patient [16, 19]. In this context, confidence measures can provide insight on the likelihood of each prediction.

2) *Unknown time to conversion.* The early prognostic of AD is very important to treat patients as well as possible and manage their expectation regarding disease progression. Furthermore, knowing how fast will the progression be in case it occurs is of great value. With such information, clinicians could timely adjust treatments and medical appointments to the need of that specific patient. In this context, we should model conversion to dementia within specific time intervals. Besides being more informative from a clinical point of view, the few studies [2, 7, 17] that addressed this question showed that prognostic models learned with patients having similar time to conversion are more reliable than those models learned from heterogeneous groups of patients (regarding their conversion time).

In this work, we propose a supervised learning approach aiming to tackle these issues. Following a stepwise learning process, this approach not only predicts conversion to dementia, but outputs the more likely window of occurrence (short-term or long-term conversion). To test this approach, we used Conformal Prediction (CP) [20, 23], a machine learning technique that produces confidence measures. Conformal Predictors (CPs) are built on top of traditional machine learning algorithms (denoted as underlying algorithms), predicting the class that makes the new example (patient) more "conform" to the training set within certain levels of confidence. A confidence level of 0.9, for instance, means that the conformal predictors commit to a maximum of 10% of errors. CPs have been applied successfully in disease-related problems such as the early detection of ovarian cancer [5], diagnosis of acute abdominal pain [15] or stroke risk assessment [10].

Neuropsychological data have proved their relevance on predicting converting MCI patients, being considered as accurate as the more complex models involving data integration [12]. Despite the value of biological biomarkers, they retain a supportive role to the neuropsychological assessment. On the one hand, measurable cognitive impairment is still a hallmark for the diagnosis of dementia and mild cognitive impairment [6, 24]. On the other hand, neuropsychological tests (NPTs) are less expensive, non-invasive, and easily applied in clinical practice.

Machine learning approaches are gaining a lot of relevance in dementia research [2, 13], but mainly focusing on brain imaging data (Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET)). Works involving only NPTs tend to use traditional statistical analysis, which may not be suitable to successfully capture all its predictive power [1].

In this work, we study the feasibility of the proposed approach to the prognostic problem of MCI-to-AD conversion within certain time windows, using the conformal prediction framework and neuropsychological data. To our knowledge, this was not explored to date. The proposed supervised learning approach may, however, be used with other methods and/or data types.

2. DATASET AND METHODS

2.1 ADNI data

The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. Its goals is to find relevant biomarkers in all stages of AD to guide future clinical trials for new possible treatments. Currently, it includes 1400 MCI and AD patients, as well as normal subjects, being followed. ADNI includes several biomarkers of Alzheimer's disease such as demographic data, neuropsychological tests, Cerebrospinal fluid (CSF), structural Magnetic Resonance Imaging, functional-MRI (fMRI), Positron Emission Tomography and other biological data. This data is collected from every ADNI participant at the baseline assessment, as well as on their annual follow-up consultations.

In this work, we used NPTs data from ADNI-2 patients. A list of the 93 features used (original or created from the data) is available at: https://fenix.tecnico.ulisboa.pt/homepage/ist165127/support-data-for-publications. This data also includes demographic data such as age, gender, education level. NPTs include, but are not limited to, the Mini Mental State Examination (MMSE) and the Alzheimer's Disease Assessment Scale – cognitive subscale (ADAS-Cog). For a detailed description of ADNI's data, we refer to http://adni.loni.usc.edu/.

The total number of patients was filtered to include only MCI patients at the baseline assessment. Reverters – patients that converted from MCI to normal cognition, or even AD to MCI, were excluded from the pool. This is usually the method employed by most studies, as reversion is clearly unexpected and can be related to errors in diagnosis or the presence of diseases other than AD [8], which affects the diagnosis. Also, patients were excluded if they didn't have any follow-up assessments.

We followed the strategy to create learning examples using time windows described in [17]. For a given time-window, we considered patients that converted to dementia within a predefined interval, i.e. which had the diagnosis of AD in one of the yearly assessments up until the limit of the window. Those are labeled cMCI (converter MCI). On the other hand, patients that didn't convert to AD during that period and presented a diagnosis of MCI at the limit of the window or afterwards, are included in the learning set labelled as sMCI (stable MCI). In this work, and taking into account the nature of the ADNI data, we chose a follow-up time of 4 years. From a total of 265 MCI patients, 143 (54%) patients remained stable and 122 (46%) converted to dementia within the follow-up period. Regarding the converting patients, 89 converted to AD within the first 2 years (short-term conversion, *s.t.* cMCI), and 33 converted between 2 and 4 years (long-term conversion, *l.t.* cMCI).

2.2 Conformal Prediction

We introduce the idea behind the conformal prediction framework. For a more formal description we refer to [23]. Let us assume that we are given a training set $\{(x_1, y_1),...,(x_{n-1}, y_{n-1})\}$, where $x_i \in X$ is a vector of attributes and $y_i \in Y$ is the class label (assuming a binary classification problem). Given a new test example (x_n) we aim to predict its class. Intuitively, we assign each class $y_n \in Y$ to x_n, at a time, and then evaluate how "non-conform" the example (x_n, y_n) is in comparison with the training data. The most likely class label conforms better with the training set. A non-conformity measure, to assess the non-conformity of the test example, must be extracted from the underlying classifier (any classifier may be used). To evaluate how different x_n is from the training set, we compare its non-conformity score with those of the remaining training examples x_j , $j = 1, \dots, n-1$, using the *p*-value function (distinct from the *p*-value from statistics):

$$p(\alpha_n) = \frac{|\{j=1,...,n\colon \alpha_j \geq \alpha_n\}|}{n} \qquad (1)$$

where α_n is the non-conformity score of x_n, assuming it is assigned to the class label y_n. If the *p*-value is small, then the test example (x_n, y_n) is non-conforming since few examples (x_i, y_i) had a higher non-conformity score when compared with α_n. On the other hand, if the *p*-value is large, x_n is very conforming since most of the examples (x_i, y_i) had a higher non-conformity score when compared with α_n. Once p-values are computed, CP can be used in one of the following ways:

1) *Using prediction regions.* For a given significance level (ε), CPs output a prediction region, T^{ε}: set of all classes with $p(\alpha_n) > \varepsilon$, contrarily to the single predictions given by standard classifiers. These prediction regions have a guaranteed error rate. This means that the frequency of errors (fraction of true values outside T^{ε}) does not exceed ε, at a confidence level $1 - \varepsilon$. The error rate is guaranteed under the randomness assumption, which states that the examples are independently drawn from the same distribution (this property is called validity) [23]. Prediction regions may therefore comprise more than one class (uncertain prediction), no class (empty prediction) or a single class (certain prediction). Multiple predictions are not errors but a reflection of the fact that the classifier is not being confident enough to predict a certain class. The smaller the prediction region the more efficient the conformal predictor is [23].

2) *Using forced predictions.* If one would rather have single predictions than prediction regions, CPs predict the class with the highest *p*-value (forced prediction), alongside with its credibility (the largest p-value) and confidence (complement to 1 of the second highest *p*-value). Confidence reveals how likely the predicted classification is compared with the other classes. Credibility reveals how suitable the CP is for classifying the given example. Low credibility means that either the training set is non-random or the test example is not representative of the training set, and thus, the predicted class is non-conforming to the training data. Given that the data was generated independently from the same distribution (randomness assumption [23]), the probability that the credibility is less than some threshold ε is less than ε [22]. The higher the values of both confidence and credibility the more reliable is the prediction.

2.2.1 Transductive and Inductive CP

Conformal prediction may be used in the transductive or in the inductive setting. When transductive framework is used, the training set is enriched with the test example, and the underlying classifier is updated. Non-conformity scores are then computed for all the training examples. This process is repeated for all class labels $y \in Y$. A new prediction is therefore based on all the training examples. For large datasets, this is computationally very demanding. This led to the emergence of inductive learning [14, 22]. When inductive is used, the training set $\{(x_1, y_1),\ldots,(x_{n-1}, y_{n-1})\}$ is divided into the proper training set $\{(x_1, y_1),\ldots,(x_m, y_m)\}$ and the calibration set $\{(x_{m+1}, y_{m+1}),\ldots,(x_{n-1}, y_{n-1})\}$, where $m < n - 1$. The proper training set is used to derive the prediction rule, by training the underlying classifier. This prediction rule is then used to classify the examples of the calibration set and the test example. Non-conformity scores are only computed with the examples of the calibration set.

2.2.2 Mondrian CP

Mondrian conformal prediction is a variant of CP that deals with imbalanced datasets [23]. When the number of examples of a given class is significantly larger than those of the other class, most errors are putatively from the minority class, limiting the applicability of these predictions. Mondrian conformal prediction applies CPs separately to each label class. The p-value is thus computed by comparing the non-conformity score of the test example against only training examples of the same class as the current hypothesis y_n:

$$p(\alpha_n) = \frac{|\{j=1,\ldots,n:\, y_j = y_n \ and \ \alpha_j \geq \alpha_n\}|}{|\{j=1,\ldots,n:\, y_j = y_n\}|} \qquad (2)$$

2.3 Stepwise learning with time windows

The supervised learning approach proposed in this work is described in Figure 1. It consists on a two-step supervised learning approach which starts by predicting conversion from MCI to AD, within a given level of confidence, and then complements it with the prediction of the most likely time window of conversion (short-term or long-term conversion). More specifically, in the first step, Model 1 predicts whether a given MCI patient will convert to dementia (*cMCI*) or remains MCI during the follow-up period (*sMCI*). A measure reflecting the confidence on the predicted class is outputted. If this value is below a certain (predefined) threshold, we consider that prediction as unpredictable (*No prediction*). If the prediction is trustworthy (above the confidence level) the prognostic is made. A trustworthy prediction of conversion (*cMCI)* is fed into Model 2, in the second step of the approach. This model predicts whether the patient will have a short-term or long-term conversion. Once again, low confident predictions are considered as unpredictable.

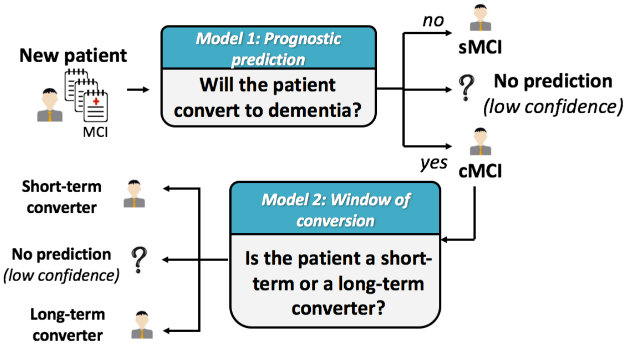
New patient
MCI
Model 1: Prognostic prediction
Will the patient convert to dementia?
no
sMCI
No prediction
(low confidence)
yes
cMCI
Short-term converter
No prediction
(low confidence)
Long-term converter
Model 2: Window of conversion
Is the patient a short-term or a long-term converter?

Figure 1. Workflow of the proposed stepwise supervised learning approach. Model 1 predicts whether a MCI patient is going to convert (cMCI: converter MCI), remains MCI (sMCI: stable MCI), or if no prognostic is presented because the prediction is below the chosen confidence level. Model 2 predicts the more likely window of conversion (short-term or long-term) for converting patients, or no prediction if it is not trustworthy.

2.3.1 Classification setup

The proposed classification approach is divided into two phases: *i)* training and tuning the parameters to Model 1 (prognostic prediction, Figure 1) and Model 2 (window of conversion, Figure 1) and *ii)* applying the stepwise approach described before: for a given test set, first classify patients as sMCI or cMCI and then in short-term (cMCI at 0-2 years) or long-term (cMCI at 2-4 years) conversion. The dataset was randomly split (keeping class proportions) in training set (80%) and test set (20%). This process was repeated 5 times with fold randomization. During the training and tuning parameters phase, for each training set, a further 5-fold cross-validation (CV) procedure was repeated 5 times with fold randomization. This aimed at dividing the examples into training and validation sets, in order to use the validation examples to determine the parameters that optimize a predefined evaluation metric. We optimized the values of F-measure and outputted also the values of sensitivity and specificity. We used Naïve Bayes and tested three parameters: 1) use Gaussian distribution or 2) kernel estimator for numerical attributes, and 3) whether to use supervised discretization to convert numerical into nominal attributes.

Since the use of preprocessing techniques to deal with a large number of (possibly irrelevant) features or imbalanced classes may have a significant impact on both classification performance and model simplification and interpretability, the worth of using/not using feature selection (FS) and/or using/not using SMOTE to deal with class imbalance was tested.

Four methods from the filter family of FS were tested. A filter feature selection algorithm evaluates the value of a features' subset without taking into account the learning algorithm that is applied afterwards. These methods take different measures in search for the most predictive features, such as Weka's *CfsSubsetEval* is an implementation of Hall's work [9], that measures the predictive power of each feature while minimizing the redundancy between them. *CorrelationAttributeEval* measures the Person's correlation between features and the class and *SymmetricalUncertAttributeEval* and *InfoGain* measure the worth of a particular attribute by its information gain and symmetrical uncertainty with respect to the class, respectively. The latter three methods rank the features by their individual evaluations and thus the number of selected features to keep must be defined. Subset sizes of 10 to 40 features, with incremental steps of 5 features, were tested. These FS methods were implemented in Weka.

Furthermore, class imbalance was tackled with the Synthetic Minority Over-sampling Technique (SMOTE) [3]. SMOTE is an oversampling technique that generates synthetic samples from the minority class by choosing a set of similar instances and perturbing the attributes by a random amount. SMOTE percentage ranges from 0% to the equilibrium of the class proportions, in 3 steps. In order to ensure the validity of the results, FS and SMOTE were only applied to the training data within each cross-validation fold.

The worth of using FS and/or SMOTE was assessed by the Wilcoxon Signed Rank Test [4] on the averaged F-Measure across the 5×5-fold CV using IBM SPSS Statistics 24 (released version 24.0.0.0).

After the tuning phase, we trained Model 1 and Model 2 with the training data (80% of the original dataset) and the optimized parameters and then, we tested the proposed stepwise approach with the test set (20% of the original dataset). The classification approach was implemented in Java using WEKA's functionalities (version 3.8.0).

2.3.2 Conformal prediction settings

Given that the dataset under study does not have high dimensionality, we used the Transductive Conformal Prediction framework. In addition, we used Mondrian CP framework to tackle the class imbalance of the dataset under study.

We used Naïve Bayes as underlying classifier of the CP approach since, in previous work [17], it outperformed other commonly used classifiers (such as SVMs, Decision Trees and Random Forests) in the MCI-to-dementia conversion problem. The following non-conformity measure was used:

$$-\log p(y_i = c|x_i), \tag{3}$$

where p is the posterior probability estimated by Naïve Bayes.

We used the Forced Prediction approach to train and tune the parameters to Model 1 and Model 2 (first phase of the proposed approach, Section 2.3.1). The confusion matrix computed to fine-tune the parameters was built with no confidence level as it used all the (forced) predictions made. In the second phase of the proposed framework (applying the stepwise approach), both approaches of Conformal Prediction – using Prediction Regions (PR approach) or Forced Predictions (FP approach) (section 2.2) were used. In the former, certain predictions were considered as trustworthy while uncertain and empty predictions were seen as unreliable (no class predicted). Three significance levels were tested ε = {0.15, 0.20, 0.30} corresponding to the following confidence levels: cfd = {0.70, 0.75, 0.85}. In the latter, forced predictions above the predefined confidence threshold were considered as trustworthy while the remaining were disregarded. Three confidence (cfd) thresholds were tested (cfd = {0.70, 0.75, 0.80}).

Two other approaches, using Prediction Regions, were tried to compute the optimization metric in the parameters' tuning phase (first phase of the proposed approach, Section 2.3.1). Specifically, F-measure was computed: *i)* using only certain predictions (disregarding the uncertain and empty predictions) and *ii)* using the certain predictions and considering the uncertain and empty predictions as misclassified examples. However, since these approaches did not enhanced the results and due to space constraints, we do not report these results.

3. RESULTS AND DISCUSSION

The data used in this work is described in Section 2.1 and summarized in Table 1.

Table 1. Data description.

	sMCI	**cMCI**	
Model 1	143 (54%)	122 (46%)	
Model 2	-	*s.t.* cMCI	*l.t.* cMCI
		89 (73%)	33 (27%)

From the empirical experiments performed using different FS methods and size of the features' subset, we chose the *InfoGain* with 25 features. Small differences were found between the averaged F-Measure with distinct FS methods. Although good results were achieved for larger subsets of features, we decided to pursued the analysis with 25 features, as it represents a good trade-off between a good classification performance and a minor number of features used, which should be as small as possible regarding the reduced sample size.

When evaluating the worth of using FS with Model 1, we concluded that there was no statistical difference (p-value<0.291) between the results obtained with or without feature selection. Still, we decided to proceed with feature selection for the sake of model interpretability. We note that SMOTE was not applied to Model 1 since its respective class imbalance is negligible. In what concerns Model 2, using SMOTE enhanced the results (p-value=0.00), either using or not using FS. Moreover, learning the model with a reduced set of features also improved the results (p-value<0.04). As such, we pursued the analysis with FS for both Model 1 and Model 2 and with SMOTE for Model 1.

As aforementioned, *InfoGain* ranks features according to their information gain in respect to the class. As such, and while not being the main goal of this work, we briefly evaluated the highest and lowest ranked features for Model 1 and Model 2. Table 2 reports the highest and the lowest 10 features for each model. Extended tables are provided in the following link: https://fenix.tecnico.ulisboa.pt/homepage/ist165127/support-data-for-publications. ADAS-Cog total scores and memory sub-scores (word recall and delayed recall) along with the total score of the Functional Assessment Questionnaire are amongst the highest ranked features of Model 1. Different measures of the Clinical Dementia Rating are also present, including the memory sub-score but also measures of daily live activities, judgment, problem solving and orientation.

Regarding the highest ranked features of Model 2, we can see sub-scores from the Neuropsychiatric Inventory that measures several behavioral patterns such as the patient's appetite, depression, irritability, apathy and sleep. These were totally absent from the highest ranked features of Model 1. Although there are common highest ranked features for both Models (such as sub-scores of ADAS-Cog and the Clinical Dementia Ranting), there are some measures, such as total scores of ADAS-Cog, part B of the trail making test and the total score of the Functional Assessment Questionnaire that appears as a top feature for Model 1 but as one of the bottom features for Model 2. These results suggest that while some measures of delayed recall are extremely important as Alzheimer's disease biomarkers, there are marked differences in the most predictive tests and scores to classify converting patients and differentiate early from late converters. Finally, some features ranked low on both Models, such as the patient's age, education level, number of years of symptoms prior to baseline and age at retirement.

Table 3 reports the results obtained with the optimized parameters tuned within a randomized 5-fold cross validation scheme, for each train set (5 × 80% data) and for Model 1 and Model 2. As aforementioned, Model 1 predicts whether a MCI patient converts to AD (positive class, cMCI) or remains stable during the follow-up period of 4 years (negative class, sMCI). The model successfully distinguished between converting and non-converting patients, as values above 0.76 were achieved in all evaluation metrics. Non-converting patients were easier to identify than those who converted, pointed by the higher values of specificity.

Table 3. Results obtained with the optimized parameters fine-tuned within a randomized 5-fold cross validation scheme, for each train fold (5 × 80% data) and for each model 1 (prognostic prediction) and 2 (time to conversion).

	F-Measure	Sensitivity	Specificity
Model 1	0.784±0.013	0.764±0.014	0.801±0.014
Model 2	0.642±0.028	0.631±0.032	0.620±0.029

In the second step of the stepwise approach, Model 2 aims to predict the most likely window of conversion, between a short-term (up to 2 years) and long-term (2-4 years) conversion. This classification task is more challenging than the prognostic prediction learned by Model 1. In fact, the classification models are based on the performance accomplished in the neuropsychological assessment made at the baseline (first patient's assessment). A converter MCI has certainly a more accentuated deficit in the NPTs when compared with a stable MCI. Between converting patients, differences in the performance on the NPTs assessment should also be present, since some patients are closer to become demented than others, although at a smaller level, which hampers the learning task. The results corroborate this idea (Model 2, Table 3). The negative class corresponds to a short-term conversion while the positive class represents the long-term conversion. Despite the difficulty of this classification task, and the small and imbalanced set of examples used in the learning process, the performance was above the random level, with F-Measure of 64% and equilibrated values of sensitivity and specificity (63% and 62%, respectively).

Tables 4 and 5 report the results obtained with the stepwise approach ran with the test set (5 × 20% data), using the models trained with the optimized parameters, and following the Forced Prediction and Prediction Regions approaches (Section 2.3.2), respectively. We note that, since Models 1 and 2 learn distinct classes (Model 1: sMCI vs cMCI and Model 2: *short-term* cMCI vs *long-term* cMCI) we cannot present an overall evaluation metric to assess their performance (as we did in Table 3). Instead, we evaluate the number of cases that are correctly (and incorrectly) classified for each class and at each step of the stepwise approach, at different confidence thresholds. We note that the class distribution in the test set is imbalanced (it has in average 24 sMCI, 18 short-term cMCI and 6 long-term cMCI) thus hampering the results, mainly during the second step.

Model 1 has a high predictive power on classifying MCI patients as being converting or non-converting, following both approaches, with about 84% of non-converting cases correctly classified, even with the highest confidence threshold (cfd=0 .80) for the FP approach and 75% for the PR approach. About 79 to 83% (cfd = 0.70 to 0.80) of the converting cases were also correctly identified, although the number of cases with short-term conversion that were correctly classified were much higher (above 83%) than those who converted between 2 and 4 years (around 60%), for the FP approach. This may be, however, an effect of the imbalanced dataset, and not of the incapability of this class to be learned. As the confidence threshold increases, the number of unreliable predictions (number of non-classifiable cases) also increases from 0 to 6%. Still, even for high values of confidence, Model 1 achieved good classification performances and a small number of non-classifiable cases. Similar results were obtained with the PR approach. In this scenario, there is however a higher amount of untrustworthy predictions (up to 32%) but also a smaller number of wrong predictions (maximum of 12%).

A final prognostic prediction is given for non-converting patients (Model 1). Those classified as converters (about 16 short-term cMCI and 3 long-term cMCI) are then fed to Model 2, to complement their prognostic with a likely window of progression. About 64% of the short-term converting patients are correctly classified with a confidence level of 0.70 for the FP approach. On the other side, this model was not reliable on identifying long-term converters since it correctly classified at maximum average of 1.4 cases out of 3.8. Once again, the small number of long-term cMCI does not allow us to evaluate whether these results are due to the challenging learning task or a consequence of the class imbalance. Still, it is clear that this classification task is harder than that learned with Model 1. The predictions are in general less confident as the number of non-classifiable cases goes from 6 to 32%, with the growth of the confidence threshold. Comparing with the PR approach (Table 5) a minor number of cases of *s.t.* and *l.t.* conversions were predicted, mainly for higher confidence levels, but also less errors were made (11% for confidence level of 0.85). This is due to the larger amount of non-classifiable patients obtained with this approach. This leads to the idea that the PR approach is more stringent on the demanded trustiness to make a prediction than the FP approach. In fact, the FP approach outputs a larger number of correctly classified cases but also a higher number of misclassified cases, as only a reduced number of cases are considered as non-classifiable.

4. CONCLUSIONS AND FUTURE WORK

This paper presents a two-step supervised learning approach which starts by predicting conversion from MCI to AD, within a given level of confidence, and then complements it with the prediction of the most likely time window of conversion (short-term or long-term conversion), using Conformal Prediction and data from ADNI. Despite the clinical meaning and significance of the proposed approach, the dataset used in this study did not allow a proper validation of the methodology, as more examples of long-term converting patients are required.

Further work is needed to improve the methodology. In particular, we aim to test the approach with a larger data sample, either using different sources of data or using follow-up assessments to build learning examples, considering those as being a "baseline" assessment. In this case, we should guarantee that examples of the same patient are bundled together, either present in the train or in the test set. Besides that, we would like to test whether cases whose conversion is at the borderline of the time window (i.e., a patient that converts for instance at 2 years and more (or less) a few months) are introducing noise in the classification task. We will also try different cut-offs of the time windows, which may putatively be more significant clinically.

Another future work concerns the use of methods to deal with multiclass problems, where the classifier should distinguish, in one step, which patients would convert in a short-term, long-term or remain stable. Different algorithms and data types should also be tested.

We highlight the importance of introducing confidence levels associated with the predictions. Firstly, it is paramount for clinicians to know how much they can rely on the prediction made for a given patient, and the time that it takes for the conversion, in order to timely adjust the treatments. Secondly, even untrustworthy predictions might be useful as clinicians can prescribe more specific exams to deeply evaluate the neurodegeneration of these patients.

5. ACKNOWLEDGMENTS

This work was partially supported by FCT under the Neuroclinomics2 project PTDC/EEI-SII/1937/2014, research grant (SFRH/ BD/95846/2013) to TP and FF (SFRH/BD/118872/2016), INESC-ID plurianual ref. UID/CEC/50021/2013, and LASIGE Research Unit ref. UID/CEC/00408/2013.

Data collection and sharing for this project was funded by the Alzheimers 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, Alzheimers Association; Alzheimers 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 Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

6. REFERENCES

[1] Barnes, D.E. et al. 2014. A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease. *Alzheimer's & dementia : the journal of the Alzheimer's Association*. 10, 6 (2014), 646–55. DOI:https://doi.org/10.1016/j.jalz.2013.12.014.

[2] Cabral, C. et al. 2015. Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. *Computers in Biology and Medicine*. 58, (2015), 101–109. DOI:https://doi.org/10.1016/j.compbiomed.2015.01.003.

[3] Chawla, N. V et al. 2002. SMOTE : Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 16, (2002), 321–357.

[4] Demsar, J. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*. 7, (2006), 1–30.

[5] Devetyarov, D. et al. 2012. Conformal predictors in early diagnostics of ovarian and breast cancers. *Progress in Artificial Intelligence*. 1, 3 (2012), 245–257. DOI:https://doi.org/10.1007/s13748-012-0021-y.

[6] Egerházi, A. et al. 2007. Automated Neuropsychological Test Battery (CANTAB) in mild cognitive impairment and in Alzheimer's disease. *Progress in neuro-psychopharmacology & biological psychiatry*. 31, 3 (2007), 746–751. DOI:https://doi.org/10.1016/j.pnpbp.2007.01.011.

[7] Eskildsen, S.F. et al. 2013. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *NeuroImage*. 65, (2013), 511–521. DOI:https://doi.org/10.1016/j.neuroimage.2012.09.058.Prediction.

[8] Grande, G. et al. 2016. Reversible Mild Cognitive Impairment: The Role of Comorbidities at Baseline Evaluation. *Journal of Alzheimer's Disease*. 51, (2016), 57–67.

[9] Hall, M.A. 1999. *Correlation-based Feature Selection for Machine Learning*. University of Waikato.

[10] Lambrou, A. et al. 2010. Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction. *Artificial Intelligence Applications and Innovations*. (2010), 146–153. DOI:https://doi.org/10.1007/978-3-642-16239-8.

[11] Langa, K. and Levine, D. 2014. The diagnosis and management of mild cognitive impairment: a clinical review. *JAMA*. 312, 23 (2014), 2551–61. DOI:https://doi.org/10.1001/jama.2014.13806.

[12] Lee, S.J. et al. 2014. A clinical index to predict progression from mild cognitive impairment to dementia due to Alzheimer's disease. *PloS one*. 9, 12 (2014), e113535. DOI:https://doi.org/10.1371/journal.pone.0113535.

[13] Moradi, E. et al. 2014. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*. 104, (2014), 398–412. DOI:https://doi.org/10.1016/j.neuroimage.2014.10.002.

[14] Norinder, U. et al. 2014. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *Journal of Chemical Information and Modeling*. 54, (2014), 1596–1603.

[15] Papadopoulos, H. et al. 2009. Reliable diagnosis of acute abdominal pain with conformal prediction. *Engineering Intelligent Systems*. 17, 2–3 (2009), 127–137.

[16] Papadopoulos, H. 2011. Reliable probabilistic prediction for medical decision support. *IFIP Advances in Information and Communication Technology*. 364 AICT, PART 2 (2011), 265–274. DOI:https://doi.org/10.1007/978-3-642-23960-1_32.

[17] Pereira, T. et al. 2017. Predicting the progression of mild cognitive impairment to dementia using neuropsychological data: a supervised learning approach using time windows. *BMC Medical Informatics and Decision Making*. 17:110, (2017). DOI:https://doi.org/DOI 10.1186/s12911-017-0497-2.

[18] Prince, M. et al. 2015. *World Alzheimer Report 2015: The Global Impact of Dementia - An analysis of prevalence, incidence, cost and trends*.

[19] Ribeiro, M.T. et al. 2016. Why should I trust you? Explaining the predictions of any classifier. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)* (2016), 4503.

[20] Shafer, G. and Vovk, V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*. 9, (2008), 371–421.

[21] Silva, D. et al. 2013. Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting. *Journal of Alzheimer's disease : JAD*. 34, 3 (2013), 681–9. DOI:https://doi.org/10.3233/JAD-122098.

[22] Toccacheli, P. et al. 2016. Conformal Predictors for Compound Activity Prediction. *Conformal and Probabilistic Prediction with Applications*. 51–66.

[23] Vovk, V. et al. 2005. *Algorithmic Learning in a Random World*. Springer.

[24] Wolfsgruber, S. et al. 2014. The CERAD neuropsychological assessment battery total score detects and predicts alzheimer disease dementia with high diagnostic accuracy. *American Journal of Geriatric Psychiatry*. 22, 10 (2014), 1017–1028. DOI:https://doi.org/10.1016/j.jagp.2012.08.021.

Table 2. List of the 10 highest and lowest ranked subsets of features obtained with *InfoGain* method for Models 1 and 2.

Highest ranked features		**Lowest ranked features**	
Model 1	***Model 2***	***Model 1***	***Model 2***
ADAS-Cog - Delayed Word Recall	RAVLT - Trial B Total	Age at retirement	Trail Making Test - Part B
ADAS-Cog - Total Score (ADAS 13)	NPI - Appetite and eating disorders: Item score	MOCA - Letter Fluency	RAVLT - Delayed Recall
FAQ - Total Score	MMSE - Total Score	NPI - Total Score	Category Fluency (Animals) - Total Correct
ADAS-Cog - Word Recall	ADAS-Cog - Orientation Score	Education Level	Trail Making Test - Part A
ADAS-Cog - Total Score (ADAS 11)	ADAS-Cog - Word Recall	Years of symptoms prior to baseline	FAQ - Total Score
CDR - Home and Hobbies Score	NPI - Depression/Dysphoria: Item score	MOCA - Digit Span Forward	NPI - Total Score
CDR - Memory Score	ADAS-Cog - Delayed Word Recall	Gender	MOCA - Delayed Recall
RAVLT - Trial 6 Total	NPI - Irritability/Lability: Item score	Diagnostic Summary - Subjective memory complaint	MMSE - Writing
CDR - Judgment and Problem Solving Score	NPI - Anxiety: Item score	MMSE - Construction - Copy Score	Trail Making Test - Part B
RAVLT - Trial B Total	MOCA - Attention - Letters and Tapping	Age at baseline	ADAS-Cog - Total Score (ADAS 11)

Table 4. Results obtained with our stepwise approach on the test set using the models trained with the optimized parameters and following the Forced Prediction (FP) approach. We note that sMCI stands for stable MCI, cMCI: MCI, *s.t.* cMCI: short-term conversion, *l.t.* cMCI: long-term conversion and cfd: confidence threshold.

	Test set				**Model 1**						**Model 2**			
	# sMCI	**# cMCI**			**# sMCI as sMCI**	**# cMCI as cMCI**			**# Misclassified**	**Non-classifiable cases (%)**	**# cMCI as cMCI**		**# Misclassified**	**Non-classifiable cases (%)**
cfd		*Total*	*s.t. cMCI*	*l.t. cMCI*		*Total*	*s.t. cMCI*	*l.t. cMCI*			*s.t. cMCI*	*l.t. cMCI*		
0.70	28	24.0	18.0	6.0	24.2±0.8	20.0±1.9	16.2±1.1	3.8±1.0	7.8±2.3	0	10.4±3.2	1.4±1.1	8.8±2.6	14±13.9
0.75	28	24.0	18.0	6.0	24.0±0.8	20.0±1.8	16.2±1.1	3.8±1.0	7.8±1.9	0.4±0.9	10.2±2.8	1.2±1.1	8.2±1.9	18±13.5
0.80	28	24.0	18.0	6.0	23.4±1.1	19.0±1.7	15.4±0.9	3.6±1.1	7.6±3.3	6.4±4.5	7.6±2.5	1.0±1.2	5.8±1.5	32.2±11.7

Table 5. Results obtained with our stepwise approach on the test set using the models trained with the optimized parameters and following the Prediction Region (PR) approach. We note that sMCI stands for stable MCI, cMCI: MCI, *s.t.* cMCI: short-term conversion, *l.t.* cMCI: long-term conversion and cfd: confidence threshold.

	Test set				**Model 1**						**Model 2**			
	# sMCI	**# cMCI**			**# sMCI as sMCI**	**# cMCI as cMCI**			**# Misclassified**	**Non-classifiable cases (%)**	**# cMCI as cMCI**		**# Misclassified**	**Non-classifiable cases (%)**
Cfd.		*Total*	*s.t. cMCI*	*l.t. cMCI*		*Total*	*s.t. cMCI*	*l.t. cMCI*			*s.t. cMCI*	*l.t. cMCI*		
0.70	28	24.0	18.0	6.0	22.0±2.1	18.2±1.8	15.2±0.8	3.0±1.0	6.2±1.9	12.0±3.2	10.6±2.2	1.2±0.8	6.2±1.9	16.8±12.2
0.80	28	24.0	18.0	6.0	23.4±1.1	19.0±1.7	15.4±0.9	3.6±1.1	5.4±2.9	7.0±3.7	8.2±2.3	1.2±1.3	5.0±1.0	35.6±14.2
0.85	28	24.0	18.0	6.0	21.4±1.3	18.6±5.9	16.2±5.7	2.4±0.9	5.0±3.2	17.2±6.4	7.8±2.0	0.4±0.9	2.4±1.5	44.6±15.1