Predicting the Progression of Mild Cognitive Impairment Using Machine Learning: A Systematic and Quantitative Review

Manon Ansart, Stéphane Epelbaum, Giulia Bassignana, Alexandre Bône, Simona Bottani, Tiziana Cattai, Raphäel Couronné, Johann Faouzi, Igor Koval, Maxime Louis, et al.

Predicting the Progression of Mild Cognitive Impairment Using Machine Learning: A Systematic and Quantitative Review

Manon Ansart[a,b,c,d,e,*], Stéphane Epelbaum[a,b,c,d,e,f], Giulia Bassignana[a,b,c,d,e], Alexandre Bône[a,b,c,d,e], Simona Bottani[a,b,c,d,e], Tiziana Cattai[a,b,c,d,e,1], Raphaël Couronné[a,b,c,d,e], Johann Faouzi[a,b,c,d,e], Igor Koval[a,b,c,d,e], Maxime Louis[a,b,c,d,e], Elina Thibeau-Sutre[a,b,c,d,e], Junhao Wen[a,b,c,d,e], Adam Wild[a,b,c,d,e], Ninon Burgos[a,b,c,d,e], Didier Dormont[a,b,c,d,e,g], Olivier Colliot[a,b,c,d,e,f,g], Stanley Durrleman[e,a,b,c,d]

[a]*Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France*
[b]*Inserm, U 1127, F-75013, Paris, France*
[c]*CNRS, UMR 7225, F-75013, Paris, France*
[d]*Sorbonne Université, F-75013, Paris, France*
[e]*Inria, Aramis project-team, F-75013, Paris, France*
[f]*Institute of Memory and Alzheimer's Disease (IM2A), Centre of excellence of neurodegenerative disease (CoEN),National Reference Center for Rare or Early Dementias, Department of Neurology, Pitié-Salpêtrière Hospital, AP-HP, Boulevard de l'hôpital, F-75013, Paris, France*
[g]*AP-HP, Pitié-Salpêtrière hospital, Department of Neuroradiology, Paris, France*
[h]*Dept. of Information Engineering, Electronics and Telecommunication, Sapienza University of Rome, Italy*

Abstract

Context. Automatically predicting if a subject with Mild Cognitive Impairment (MCI) is going to progress to Alzheimer's disease (AD) dementia in the coming years is a relevant question regarding clinical practice and trial inclusion alike. A large number of articles have been published, with a wide range of algorithms, input variables, data sets and experimental designs. It is unclear which of these factors are determinant for the prediction, and affect the predictive performance that can be expected in clinical practice. We performed a systematic review of studies focusing on the automatic prediction of the progression of MCI to AD dementia. We systematically and statistically studied the influence of different factors on predictive performance.

Method. The review included 172 articles, 93 of which were published after 2014. 234 experiments were extracted from these articles. For each of them, we reported the used data set, the feature types (defining 10 categories), the algorithm type (defining 12 categories), performance and potential methodological issues. The impact of the features and algorithm on the performance was evaluated using t-tests on the coefficients of mixed effect linear regressions.

*Corresponding author
Email address: manon.ansart@inria.fr (Manon Ansart)

Results. We found that using cognitive, fluorodeoxyglucose-positron emission tomography or potentially electroencephalography and magnetoencephalography variables significantly improves predictive performance compared to not including them (p=0.046, 0.009 and 0.003 respectively), whereas including T1 magnetic resonance imaging, amyloid positron emission tomography or cerebrospinal fluid AD biomarkers does not show a significant effect. On the other hand, the algorithm used in the method does not have a significant impact on performance. We identified several methodological issues. Major issues, found in 23.5% of studies, include the absence of a test set, or its use for feature selection or parameter tuning. Other issues, found in 15.0% of studies, pertain to the usability of the method in clinical practice. We also highlight that short-term predictions are likely not to be better than predicting that subjects stay stable over time. Finally, we highlight possible biases in publications that tend not to publish methods with poor performance on large data sets, which may be censored as negative results.

Conclusion. Using machine learning to predict MCI to AD dementia progression is a promising and dynamic field. Among the most predictive modalities, cognitive scores are the cheapest and less invasive, as compared to imaging. The good performance they offer question the wide use of imaging for predicting diagnosis evolution, and call for further exploring fine cognitive assessments. Issues identified in the studies highlight the importance of establishing good practices and guidelines for the use of machine learning as a decision support system in clinical practice.

Keywords: quantitative review, Alzheimer's disease, Mild Cognitive Impairment, progression, automatic prediction, cognition

1. Introduction

The early diagnosis of Alzheimer's disease (AD) is crucial for patient care and treatment. Machine learning algorithms have been used to perform automatic diagnosis and predict the current clinical status at an individual level, mainly in research cohorts. Individuals suffering from mild cognitive impairment (MCI) are however likely to have a change of clinical status in the coming years, and to be diagnosed with AD or another form of dementia. Distinguishing between the MCI individuals that will remain MCI (MCI stable, or sMCI) from those who will progress to AD (pMCI) is an important task, that can allow for the early care and treatment of pMCI patients. In this article, we will review methods that have been proposed to automatically predict if an MCI patient will develop AD dementia in the future by performing a careful reading of published articles, and compare them through a quantitative analysis.

The application of machine learning to precision medicine is an emerging field, at the cross roads of different disciplines, such as computer science, radiology or neurology. Researchers working on the topic usually come from one field or the other, and therefore do not have all the skills that are necessary to design methods that would be efficient and following machine learning best practices, while being understandable and useful to clinicians.

Reviews of the automatic prediction of the current clinical diagnosis in the context of AD have already been published, but none specifically target the prediction of

progression from MCI to AD dementia. They focus on the use of magnetic resonance imaging (MRI) (Falahati et al., 2014; Leandrou et al., 2018), or of neuroimaging data more broadly (Rathore et al., 2017; Arbabshirani et al., 2017; Haller et al., 2011; Sarica et al., 2017). Several of them are systematic reviews such as Arbabshirani et al. (2017) with 112 studies on AD, Rathore et al. (2017) with 81 studies, Falahati et al. (2014) with 50 studies and Sarica et al. (2017) with 12 studies. They often gather the findings of each individual article and compare them, but no quantitative analysis of performance is proposed.

We propose here to perform a systematic and quantitative review of studies predicting the evolution of clinical diagnosis in individuals with MCI. We will report different quantitative and qualitative characteristics of the proposed method such as the sample size, type of algorithm, reported accuracy, identification of possible issues. We will then analyze this data to identify the characteristics which impact performance the most, and propose a list of recommendations to ensure that the performance is well estimated, and that the algorithm would have the best chance to be useful in clinical practice.

2. Materials and Method

2.1. Selection process

The query used to find the relevant articles was composed of 4 parts:

1. As we study the progression from MCI to AD, the words MCI and AD should be present in the abstract ;
2. We removed the articles predicting only the current diagnosis by ensuring the words "prediction" and "progression" or associated terms are present in the abstract ;
3. A performance measure should be mentioned ;
4. A machine learning algorithm or classification related key-word should be in the abstract. This fourth part ensures the selected articles make individual predictions and reduces the presence of group analyzes.

The full query can be found in Appendix A.1. Running it on Scopus on the 13^{th} of December 2018 resulted in 330 articles. The abstracts were read to remove irrelevant articles, including studies of the progression of cognitively normal individuals to MCI, automatic diagnosis methods, review articles and group analyses. After this selection 206 articles were identified. As this first selection was quite conservative, 34 additional articles were removed from the selection for similar reasons during the reading process, leaving 172 studied articles. The selection process is described in Figure S1 in Appendix A.2.

2.2. Reading process

For each study, the number of individuals was first assessed and noted. Only studies including more than 30 sMCI and 30 pMCI (111 articles) were then fully read, as we consider that experience using less than 30 individuals cannot provide robust estimates of performance. Articles with less than 30 individuals in each category were still

considered when studying the evolution of the number of articles with time, and of the number of individuals per article with time. The studies including enough individuals were then analyzed by one of the 19 readers participating in this review, and a global check was performed by one author (MA) to ensure homogeneity. 36 items, of which a list is available in Appendix A.3, were reported for each study, including the used features, the cohort, the method (time to prediction, algorithm, feature selection, feature processing), the evaluation framework and the performance measures, as well as identified biases in the method. When several experiments were available in an article, they were all reported in the table. A total of 234 experiments was thus studied.

2.3. Quality check

Several methodological issues were identified during the reading process. This list of issues was not previously defined, it has been established as issues were encountered in the various studies. We identified the following list of issues:

- Lack of a test data set: use of the same data set for training and testing the algorithm, without splitting the data set or using any kind of cross-validation method. The performance computed this way is the training performance, whereas a test performance, computed on a different set of individuals, is necessary to measure the performance that could be obtained on any other data set (i.e. generalizability of the method) .

- Automatic feature selection performed on the whole data set. When a large number of features is available, automatic feature selection can be performed in order to identify the most relevant features and use them as input. A variety of automatic algorithms exist to do this. Some studies performed this automatic feature selection on the whole data set, before splitting it into a training and a test set or performing cross-validation. An example of this issue is, first, using t-tests to identify features that best separate pMCI from sMCI, using the whole data set, then splitting the data set into a training and a test set, to respectively train the classification algorithm and evaluate its performance. In this example, the individuals from the test set have been used to perform the automatic feature selection and choose the most relevant features. This is an issue, as individuals in the test set should be used for performance evaluation only.

- Other data-leakage. More broadly, data leakage is the use of data from the test set outside of performance evaluation. Using the test data set for parameter tuning, or for choosing the best data set out of a large number of experiments, are two common examples of data leakage.

- Feature embedding performed on the whole data set. Feature embedding (for example principal components analysis) transforms the input features into a lower-dimension feature space. It is often used to reduce the input dimension when many features are available, but it does not use the individual labels (sMCI/pMCI) to do so, as feature selection often does. This issue is therefore similar to performing feature selection on the whole data set, except that only the features of the test individuals are used, and not their labels.

- Use of the date of AD diagnosis to select the input visit of pMCI individuals. An example of this issue is using the visit 3 years before progression to AD for pMCI subjects, and the first available visit for sMCI subjects, to predict the progression to AD at 3 years, even for testing the method. In this case, the date of progression to AD of the individuals of the test set was used to select the input visit, which is not possible in clinical practice, as the date of progression is not known.

Other methodological issues, not belonging to these categories, were also reported, such as incompatibility between different reported measures. The articles in which at least one of these issues was identified were not used when analyzing the performance of the methods and the method characteristics impacting them.

2.4. Statistical analysis

We studied the impact of various method characteristics (input features, algorithm...) on the performance of the classification task, separating sMCI form pMCI individuals. Several experiments were reported for each article, so we had to account for the dependency between experiments coming from the same article. In order to do so, we used linear mixed-effects models with a random effect on the article, and tested if the considered characteristics had a significant impact by performing a two-sided t-test on the corresponding regression coefficient. Only the characteristics found in more than one article with an associated performance measure were taken into account. Unless stated otherwise, the performance measure used for testing is the area under the receiver operating characteristic (ROC) curve (AUC), experiments with no reported AUC were therefore not taken into account in these tests. When testing the impact of various characteristics at the same time, conditionally to each other (e.g. among all input features, which ones have an impact on the performance when taking the other features into account), we performed a linear mixed effect regression with all these characteristics as input. Concerning the input features, d being the number of features:

$$AUC = \alpha_1 * feature_1 + ... + \alpha_d * feature_d + \beta + \beta_{article} \quad (1)$$

When testing the impact of different characteristics independently (e.g. for each algorithm, the effect of using this specific algorithm or any other), an individual linear mixed effect regression was performed for each one separately:

$$AUC = \alpha_i * algo_i + \beta + \beta_{article} \quad (2)$$

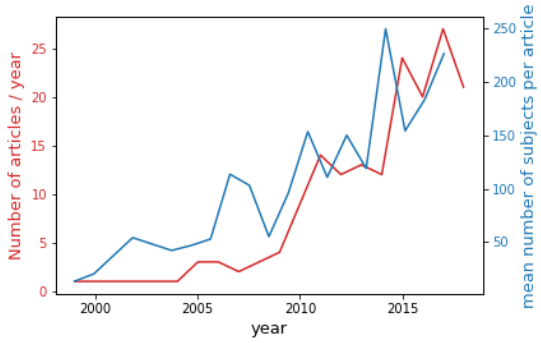
for all i, i being the algorithm number.

In both cases, a two-sided t-test was performed on α to test the significance of each coefficient. The p-values corrected for multiple comparisons were obtained by using the Benjamini-Hochberg procedure.

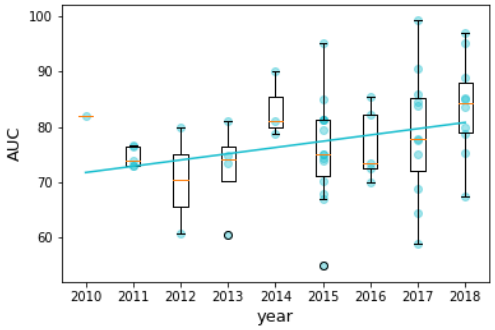
3. Descriptive analysis

3.1. A recent trend

We observe from Figure 1a that the number of articles published each year on the prediction of the progression of MCI to AD dementia has been steadily increasing since 2010.



(a) Evolution of number of article per year and of the number of individuals per article



(b) Evolution of the AUC with time

Figure 1: Recent trends. (a) Evolution of number of article per year (in red) and of the number of individuals per article with time (in blue). (b) Evolution of the area under the ROC (receiver operating characteristic) curve (AUC) with time. The AUC of each article is represented by a dot. The AUC of articles published the same year is represented as box-plots. The plain line corresponds to the regression of the AUC against time

Figure 1a also shows that the number of individuals used for the experiments is increasing over time ($p = 10^{-5}$). 84.6% of articles used data of the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Starting in 2004, this multicenter longitudinal study provides multiple modalities for the early detection of AD. As the recruitment of this largely used cohort is still ongoing, it is not surprising to see the number of included individuals increasing over the years. Studies often select individuals with a minimal follow-up time, of 3 years for example, and over the years more and more MCI individuals from the ADNI cohort fulfill these criteria, so more individuals can be included.

As shown in Figure 1b, the reported AUC are also increasing over time ($p = 0.045$), which can have multiple explanations. First, as new studies often compare their performance with those of previous methods, they tend to be published only when the obtained results seem competitive compared to previous ones. A more optimistic interpretation would be that algorithms tend to improve, and that newly available features might have a better predictive power. It has also been shown (Ansart et al., 2019; Domingos, 2012) that having a larger data set leads to a higher performance, so there may be a link between the increase in data set size and the increase in performance.

3.2. Features

T1 MRI, cognition and socio-demographic features are used in respectively 69.2%, 43.2% and 33.8% of experiments. On the other hand, fluorodeoxyglucose (FDG) positron emission tomography (PET), APOE and cerebrospinal fluid (CSF) AD biomarkers are used in 15 to 20% of experiments, and the other studied features (white matter hyper-intensities, electroencephalography (EEG), magnetoencephalography (MEG), PET amyloid, amyloid binary status without considering the PET or CSF value, diffusion tensor imaging (DTI) and PET Tau) are used in less than 10% of experiments. No study using functional MRI has been identified.

Studies using T1 MRI mainly use selected regions of interest (46.8%), whereas 34.7% use the whole brain, separated into regions of interest, and 18.5% use voxel features. Studies using neuro-psychological tests mainly use aggregated tests evaluating multiple domains of cognition (51.2% of them), and 37.4% of them combine aggregated tests with domain-specific ones. Seven experiments use new or home-made cognitive tests. 35.7% of experiments use only T1 MRI (apart from socio-demographic features), and 15% use cognition only.

The prevalence of T1 MRI does not seem surprising, as researchers working on automatic diagnosis often come from the medical imaging community, and T1 MRI is the most widely available modality. The prevalence of the imaging community can also explain the choice of cognitive features, and why more detailed and targeted cognitive tests are not used as much as more general and more well-known ones.

3.3. Algorithm

Support vector machines (SVM) and logistic regressions are the most used algorithms, being used in respectively 34.5% and 15.0% of experiments. Other algorithms are used in less than 10% of cases. Figure 2 shows the evolution of the algorithm use over time.

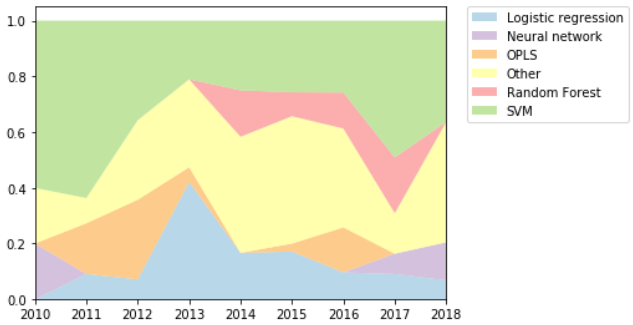


Figure 2: Evolution of the use of various algorithms with time. OPLS: orthogonal partial least square; SVM: support vector machine

The high proportion of methods using an SVM has already been shown for the prediction of the current diagnosis in Falahati et al. (2014) and Rathore et al. (2017), it is therefore not surprising that this algorithm is also commonly used for the prediction of future diagnosis. The predominance of SVM and logistic regression still seems surprising, as more recent algorithms are more popular nowadays. We see for example that random forests started being used around 2014, but the proportion of methods using this algorithm, even recently, stays low compared to the proportion of methods using an SVM. Neural networks started being used during the last two years, as it can be seen in Figure 2, and we can assume the phenomenon has been too recent to be visible just yet in the field. Overall, even if the proportion of SVM has been decreasing until 2013, the field has not been so prompt to use new algorithms as one could have expected.

3.4. Validation method

For evaluating their performance, 29.1 % of experiments use a 10-fold, and 12.8% use a k-fold with k different from 10. Leave-one individual out is also quite popular, being used in 17.5% of cases. We noted that 7.3% of experiments were trained and tested on the same individuals, and 7.3% train the method on a first cohort and test it on a different one.

It should be kept in mind when comparing the performance of different studies that the cross-validation methods can impact the performance. Using a larger training set and smaller test set is more favorable, hence the same method might result in a better performance when evaluated using a leave-one out validation than using a 10-fold validation, as shown in Lin et al. (2018). Bias and variance also vary across validation methods (Efron, 1983).

4. Performance analyses

4.1. Features

We measured the impact on the AUC of each feature compared to the others by using a linear mixed-effect model including all features used in more than one article.

The results are presented in the first part of Table 1, and show that the performance is significantly better when using cognition (p = 0.046), FDG PET (p=0.009) or EEG and MEG (p=0.003).

We also considered the impact of using each feature alone compared to a combination of them, by testing each feature independently using a linear mixed effect regression. We only tested the impact of the features that were used alone (or in combination with socio-demographic features) more than once with an associated AUC, and that had been combined with other features more than once, that is T1 MRI, cognition, and FDG PET. It is significantly better to combine T1 MRI with other features than to use it solely (p = 0.009, coefficient = 5.5). The effect is not significant for cognition (p=0.19, coefficient=3.0) and FDG PET (p=0.38 , coefficient = -6.1).

We distinguished between global neuro-psychological tests, domain-targeted tests and newly proposed tests. We measured the impact of the type of test on the AUC by performing independent regressions for each category. Experiences using a domain-specific test had a significantly greater AUC than those that did not (p=0.023, coefficient = 5.0), whereas the effect was not significant for the other two categories (p > 0.1). We tested the impact on the AUC of using longitudinal data (repeated visits as input), and of combining images of different modalities, and both were not significant (p > 0.2)

4.2. Cognition

Cognitive variables can be easily collected in clinical routine, at a low cost, and they are proven to increase the performance of the methods, so their use should be encouraged. This finding is consistent with comparisons performed in several studies. Minhas et al. (2018); Kauppi et al. (2018); Ardekani et al. (2017); Tong et al. (2017); Gavidia-Bovadilla et al. (2017); Moradi et al. (2015); Hall et al. (2015); Fleisher et al. (2008) showed that using cognition and T1 MRI performed better than using T1 MRI only. Dukart et al. (2015); Cui et al. (2011); Thung et al. (2018); Li et al. (2018) showed that adding cognition to other modalities also improved the results.

More surprisingly, we showed that using other modalities does not significantly improve the results compared to using cognition only. Although Fleisher et al. (2008) shows that using T1 MRI in addition to cognition does not improve the performance compared to using cognition only, several studies show the opposite on various modalities (Samper-Gonzalez et al., 2019; Moradi et al., 2015; Ardekani et al., 2017; Li et al., 2018; Kauppi et al., 2018). However, when taking all studies into account, it appears that the improvement one gains by including other modalities along with cognitive variables is not significant. As the cost of collecting cognitive variables compared to performing an MRI or a FDG PET is quite low, the non-significant improvement in performance might not be worth the cost and logistics of collecting data from other modalities specifically to address this question. Methods focusing on cognition only, such as proposed by Johnson et al. (2014), should therefore be further explored. Such methods should include domain-specific cognitive scores, which have shown to increase the performance.

4.3. Medical imaging and biomarkers

Imaging modalities are not as widely available as cognitive feature, but they can represent a good opportunity to better understand the disease process by showing the changes that appear before the individuals progress to AD dementia. Among the used imaging modalities, we showed that using FDG PET leads to a better performance. Similar observations have been made by Samper-Gonzalez et al. (2018). PET images could therefore represent a better alternative for the imaging community than T1 MRI, which does not significantly improve the results and should not be used alone as it leads to lower results. Changes in FDG PET appear earlier in the AD process than changes in structural MRI (Jack et al., 2010), therefore these changes might already be visible in MCI individuals several years before their progression to AD, which can explain why FDG PET is more predictive of this progression.

No method using Tau PET has been identified in this review. This new modality should also be affected early in the disease process, and could therefore represent great hopes for the imaging community. However, surprisingly, Amyloid PET or CSF value, which is also one of the earliest markers, did not have a significant impact on the prediction performance.

The use of EEG or MEG had a significant impact on the performance. However, only six experiments use these features, it is therefore difficult to conclude if this effect is real, and if it is not due to methodological issues that have not been identified during the quality check.

4.4. Combination of different imaging modalities

Multimodality has been put forward in the reviews of AD classification (Rathore et al., 2017; Falahati et al., 2014; Arbabshirani et al., 2017). As different imaging modalities correspond to various stages of the AD process, combining them could give a more complete overview of each individual. However, we did not find the impact of the use of multimodality to be significant. This result is not surprising, as the most combined modalities are MRI and FDG PET (19 out of 35 experiments using multimodality), and we showed that including other features does not lead to a significant increase in performance compared to using FDG PET alone. In addition, the cost of collecting images of different modalities for each patient is not small, and should not be neglected when using such approaches.

4.5. Longitudinal data

In a similar manner, longitudinal data could give a better view of the evolution of the patient, and hence be more predictive of the progression to AD than cross-sectional data. Nonetheless, we did not find the use of longitudinal data to have a significant effect on the performance. Similar findings are reported in Aksman (2017) for the classification of AD and in Schuster et al. (2015) for progressive diseases in general.

4.6. Algorithms

We studied the impact of the algorithms on the AUC, by using an independent linear mixed effect model on each algorithm. The results, presented in the second part of

Feature or algorithm	**coeff.**	**p-value**	**corrected p-value**	**number of exp.**
T1 MRI	2.217	0.22	0.38	103
Neuro-psychological tests	3.934	**0.046**	0.11	64
socio-demographic	0.652	0.83	0.83	59
APOE	4.612	0.092	0.18	35
FDG PET	6.768	**0.0092**	**0.037**	29
CSF	2.232	0.38	0.41	26
Others	3.12	0.28	0.4	18
EEG/MEG	16.573	**0.0025**	**0.015**	6
PET Amyloid	7.743	0.3	0.4	6
White matter hyper-intensities	-5.18	0.36	0.41	5
SVM	-4.8	0.061	0.24	35
Logistic regression	0.8	0.812	0.93	15
Random Forest	4.1	0.166	0.5	13
MKL	-0.3	0.950	0.95	10
Other	0.8	0.851	0.93	7
Bayes	5.4	0.271	0.65	6
Linear regression	-5.2	0.434	0.74	6
Neural network	10.1	**0.010**	0.06	6
OPLS	-15.5	**0.003**	**0.04**	6
Survival analysis	2.0	0.810	0.93	6
Threshold	1.1	0.791	0.93	6
LDA	-6.3	0.325	0.65	5

Table 1: Impact of features and algorithm. Benjamini-Hochberg procedure was applied to get corrected p-values. coeff.:coefficient, such as defined in Equations 1 and 2; MRI: magnetic resonance imaging; APOE: Apolipoprotein E; FDG: fluorodeoxyglucose; PET: positron emission tomography; CSF: cerebrospinal fluid; EEG: electroencephalography; MEG: magnetoencephalography; LDA: linear discriminant analysis; MKL: multiple kernel learning; OPLS: orthogonal partial least square; SVM: support vector machine

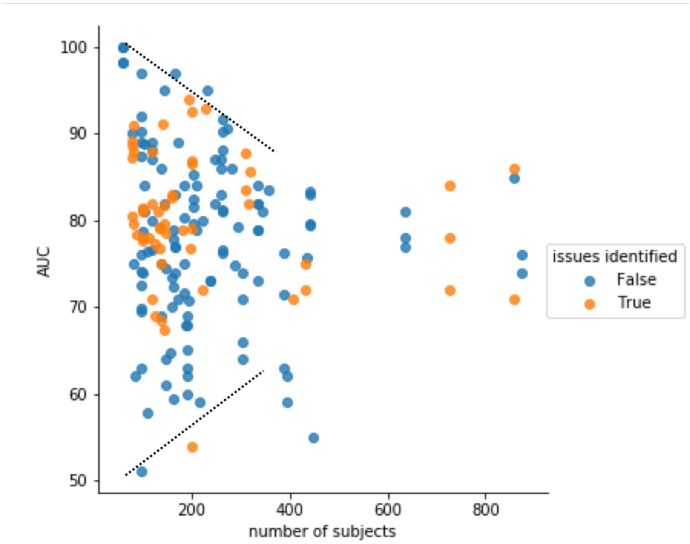


Figure 3: Relationship between the AUC (area under the ROC curve) and the number of individuals. The black dotted lines represent the upper and lower limits.

Table 1, show that the orthogonal partial least square (OPLS) algorithm performs significantly worse than others (p=0.003), whereas neural networks perform significantly better (p=0.01).

Only six experiments have been performed using each of these algorithms, so an unidentified methodological issue in one of them could greatly impact these results. As neural networks have a large number of parameters, which are often tuned manually using the test error, we found that experiments using this algorithm have high proportion of data leakage. This is consistent with the findings of Wen et al. (2019), a literature review conducted on the use of deep learning for AD classification. No conclusion regarding the impact of the classification algorithm can therefore be drawn from our results, which might be explained by the variety of algorithms, and hence the small sample size for each of them.

5. Design of the decision support system and methodological issues

5.1. Identified issues

5.1.1. Lack or misuse of test data

The lack of a test data set is observed in 7.3% of experiments. In 16% of articles using feature selection, it is performed on the whole data set, and 8% of articles do not describe this step well enough to draw conclusions. Other data leakage (use of the test set for decision making) is identified in 8% of experiments, and is unclear for 4%.

Overall, 26.5% of articles use the test set in the training process, to train the algorithm, choose the features or tune the parameters. This issue, and in particular perform-

ing feature selection on the whole data set, has also been pointed out by Arbabshirani et al. (2017) in the context of brain disorder prediction.

5.1.2. Performance as a function of data set size

We plot the AUC against the number of individuals for each experiment in Figure 3, with the colored dots representing experiments with identified issues. The colored dots show that there is a higher prevalence of studies with identified issues among high-performance studies: a methodological issue has been identified in 18.5% of experiments with an AUC below 75%, whereas this proportion rises to 36.4% for experiments with an AUC of 75% or higher (significant difference, with p = 0.006). We can observe an upper-limit (shown in dashed line) decreasing when the number of individuals increases, suggesting that high-performance achieved with a small number of subjects might be due to over-fitting. This phenomenon has already been identified by Arbabshirani et al. (2017). A lower limit is also visible, with the AUC increasing with the number of individuals. This may reflect the fact that, on average, methods generalize better when correctly trained on larger data sets. But it might also suggest that it is harder to publish a method with a relatively low performance if it has been trained on a large number of subjects, such a paper being then considered as reporting a negative result. Within papers also, authors tend to focus on their best performing method, and rarely explain what they learned to achieve this. As the number of subjects increases, the two lines seem to converge to an AUC of about 75%, which might represent the true performance for current state-of-the-art methods.

Figure 3 seems to highlight possible unconscious biases in the publications of scientific results in this field. It might be considered more acceptable to publish high-performance methods with small sample size than a low-performance method with large sample size. First, we think that low-performance methods trained on large sample size should be published also, as it is important for the field to understand what works and also what does not. In particular, we think that we, as authors, should not only focus on our best performing method, but report also other attempts. Second, it might not be such a problem that innovative methodological works that do not result in a higher performance are published also, provided that the prediction performance is not used to argue about the interest and validity of the method. The machine learning field has the chance to have simple metrics, such as AUC or accuracy, to compare different methods on an objective basis. However, we believe that one should use such metrics wisely not to discourage the publication of innovative methodological works even if it does not yield immediately better prediction performance, and not to overshadow the need to better understand why some methods work better than others.

5.1.3. Use of features of test subjects

Feature embedding is performed on the whole data set in 6.8% of experiments, meaning that the features of the test individuals are used for feature embedding during the training phase. As the diagnosis of the test individuals is often not used for feature embedding, as it is for feature selection, performing it on test individual can be considered a less serious issue than for feature selection. It however requires to re-train the algorithm each time the prediction has to be made on a new individual, which is not suited for a use in clinical practice.

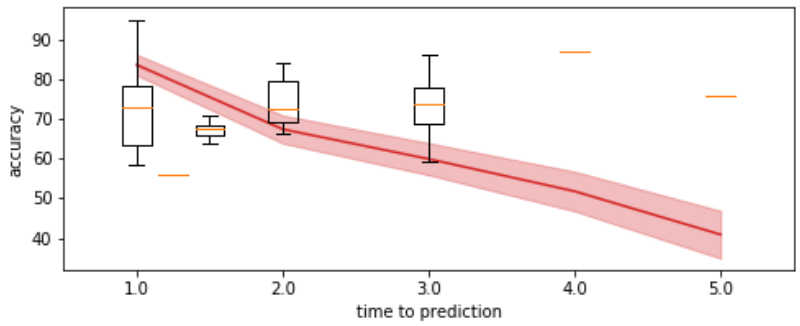


Figure 4: Evolution of the performance with respect to the time to prediction. Box plots represent the accuracy of the articles using ADNI. The straight line represents the accuracy of predicting that all individuals remain MCI, corresponding to the proportion of MCI individuals in ADNI staying MCI at the corresponding follow-up time. The shaded area corresponds to the 90% confidence interval of this percentage.

5.1.4. Use of the diagnosis date

In 5.6% of the experiments, the date of AD diagnosis is used to select the input visit of pMCI individuals, for training and testing. As explained in section 2.3, this practice can prevent the generalization of the method to the clinical practice, as the progression date of test individuals is by definition unknown.

These type of experiments answer the question "may one detect some characteristics in the data of a MCI patient 3 years before the diagnosis which, at the same time, is rarely present in stable MCI subjects?". Which should not be confused with: "can such characteristics predict that a MCI patient will progress to AD within the next 3 years". What misses to conclude about the predictive ability is to consider the MCI subjects who have the found characteristics and count the proportion of them who will not develop AD within 3 years.

This confusion typically occurred after the publication of Ding et al. (2018). The paper attracted a great attention from general media, including a post on Fox News (Wooller, 2018), stating “Artificial intelligence can predict Alzheimer’s 6 years earlier than medics”. However, the authors state in the paper that “final clinical diagnosis after all follow-up examinations was used as the ground truth label”, thus without any control of the follow-up periods that vary across subjects. Therefore, a patient may be considered as a true negative in this study, namely as a true stable MCI subject, whereas this subject may have been followed for less than 6 years. There is no guarantee that this subject is not in fact a false negative for the prediction of diagnosis at 6 years.

5.1.5. Choice of time-to-prediction

We find that 22.6% of experiments work on separating pMCI from sMCI, regardless of their time to progression to dementia. We advise against this practice, as the temporal horizon at which the individuals are likely to progress is an important information in clinical practice. Methods predicting the exact progression dates, such as what is asked in the Tadpole challenge (Marinescu et al., 2018), should be favored over methods predicting the diagnosis at a given date.

The other experiments have set a specific time to prediction, often between 1 and 3 years, meaning that they intend to predict the diagnosis of the individual at the end of this time interval. Figure 4 shows the evolution of the accuracy of these methods tested on ADNI with respect to the time to prediction. The time to prediction did not have a significant effect on AUC, accuracy, balanced accuracy, specificity nor sensitivity. Figure 4 also shows the accuracy that one would get on ADNI when using a constant prediction, that is predicting that all individuals stay MCI at future time points. The accuracy of this constant prediction has been computed using the proportion of MCI remaining stable at each visit. We show that most methods predicting the progression to AD within a short-term period smaller of 3 years do not perform better than this constant prediction. We therefore advise to use a time to prediction of at least 3 years, as for a shorter time interval the proportion of MCI individuals progressing to AD is small, predicting that all individuals remain stable therefore gives a better accuracy than most proposed methods.

This fact also shows that the accuracy may be arbitrarily increased by using a cohort with a large proportion of stable subjects. The algorithm may then yield high accuracy by mimicking a constant predictor. This effect may be alleviated by optimizing the balanced accuracy instead of the accuracy.

5.1.6. Problem formulation and data set choice

A common theme that arises from the previous issues is that the methods are not always designed to be the most useful in clinical practice. It is for example true of methods that do not use a specific time-to-prediction, or that use the date of AD diagnosis to select the included visits.

More generally, we think the most useful decision support system should not only focus on Alzheimer's disease but perform differential diagnosis. Clinicians do not usually need to distinguish between individuals who will develop AD and individuals who will not develop any neurological disorder. They most likely need help to determine which disorder an MCI individual is likely to develop. Unfortunately, no widely available data set allows the development methods for differential diagnosis to date. Methods focusing on AD should therefore target individuals who have already been identified as at risk of developing AD, by providing insight on the date at which this conversion is likely to happen. Such methods could be trained on MCI subjects that are at risk to develop Alzheimer's disease, defined for instance as the ones who have a MMSE of 27 or smaller and are amyloid positive. In addition to being closer to what can be expected in clinical practice, such data sets of at risk subjects should include a larger proportion of pMCI, leading to a better performance compared to the constant prediction. For example in ADNI, 71.6% of MCI subjects stay stable 2 years after inclusion, whereas this proportion drops to 53.7% for MCI subjects who are amyloid positive and have a MMSE of 27 or lower.

The diagnosis of Alzheimer's disease highly depends on the clinical practice, and varies greatly across sites and countries (Beach et al., 2012). Therefore, the short-term prediction of progression to Alzheimer's disease are unlikely to generalize well outside the well controlled environment of a research study. An interesting alternative may be to predict the changes in the imaging or clinical biomarkers in time rather the change in diagnosis, such as proposed by Koval et al. (2018) and Iddi et al. (2019).

5.2. Proposed guidelines

In order to ensure that the proposed method is useful for clinical practice and that the evaluated performance reflects what could be expected in real life, we propose a list of attention points:

- Separate train and test data sets by using independent cohorts or, if not available, cross-validation.
- No element of the test data set, both labels and features, should be used except for performance evaluation.
- Always pre-register the time window within which one aims to predict conversion to AD, or predict the date of progression.
- Use a large data set or pool different cohorts to obtain a large data set.
- Define a cohort that best reflects the future use of the method in clinical practice. For instance, select subjects that will be considered as at risk of developing the disease rather than all possible ADNI subjects.
- Systematically benchmark the method against the prediction that the subjects will remain stable over time.

6. Conclusion

We conducted a systematic and quantitative review on the automatic prediction of the evolution of clinical status of MCI individuals. We reported results from 234 experiments coming from 111 articles. We showed that studies using cognitive variables or FDG PET reported significantly better results than studies that did not. These modalities should be further explored, cognition because it can be easily collected in clinical routine, and FDG PET for the interest it might represent for the imaging community and for increasing our understanding of the disease. On the other hand, we showed that using solely T1 MRI yields a significantly lower performance, despite the great number of methods developed for this imaging modality. These findings call into question the role of imaging, and more particularly of MRI, for the prediction of the progression of MCI individuals to dementia. It would therefore be interesting to shift our focus towards other modalities. More specific cognitive tests could be created, and the impact of using digitized tests, that can be frequently used at home by the patients themselves, should be studied.

We identified several key points that should be checked when creating a method which aims at being used as a clinical decision support. When possible, an independent test set should be used to evaluate the performance of the method, otherwise a test set can be separated by carefully splitting the cohort. In any case, the test individuals should not be used to make decisions regarding the method, such as the selection of the features or parameter tuning. The time window in which one aims at predicting the progression to AD should be pre-registered, as the temporal horizon at which an individual is likely to progress to AD is a useful information for clinicians. Alzheimer's

disease being a very slowly progressive disease, algorithm performance should be systematically compared with the prediction that no change will occur in the future. We have shown indeed that the constant prediction may yield very high performance depending on the time frame of the prediction and the composition of the cohort. Finally, the cohort on which the method is tested should be carefully chosen and defined, so as to reflect the future use in clinical practice as best as possible. At a time where one has great expectation regarding the use of artificial intelligence to support the development of precision medicine, it becomes urgent that the field of AD research adopts state-of-the-art standards and good practices in machine learning.

Acknowledgements

Federica Cacciamani, Baptiste Couvy-Duchesne, Pascal Lu and Wen Wei participated in reading articles to conduct this review.

The research leading to these results has received funding from the program "Investissements d'avenir" ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6) from the European Union H2020 program (project EuroPOND, grant number 666992, project HBP SGA1 grant number 720270), from the ICM Big Brain Theory Program (project DYNAMO, project PredictICD), from the Inria Project Lab Program (project Neuromarkers), from the European Research Council (to Dr Durrleman project LEASP, grant number 678304), from the Abeona Foundation (project Brain@Scale). OC is supported by a "contrat d'interface local" from AP-HP. China Scholarship Council supports J.W's work on this topic.

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

References

Aksman, L.M., 2017. Longitudinal neuroimaging features for discriminating early neurodegeneration. Ph.D. thesis. King's College London.

Ansart, M., Epelbaum, S., Gagliardi, G., Colliot, O., Dormont, D., Dubois, B., Hampel, H., Durrleman, S., for the Alzheimer's Disease Neuroimaging Initiative* and the INSIGHT-preAD study, 2019. Reduction of recruitment costs in preclinical AD trials: validation of automatic pre-screening algorithm for brain amyloidosis. Statistical Methods in Medical Research , 096228021882303 6doi:10.1177/0962280218823036.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. NeuroImage 145, 137–165. doi:10.1016/j.neuroimage.2016.02.079.

Ardekani, B.A., Bermudez, E., Mubeen, A.M., Bachman, A.H., 2017. Prediction of Incipient Alzheimer's Disease Dementia in Patients with Mild Cognitive Impairment. Journal of Alzheimer's Disease 55, 269–281. doi:10.3233/JAD-160594.

Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W., 2012. Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer's Disease Centers, 2005–2010. Journal of Neuropathology and Experimental Neurology 71, 266–273. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3331862/, doi:10.1097/NEN.0b013e31824b211b.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T., Jin, J.S., Initiative, t.A.D.N., 2011. Identification of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Multivariate Predictors. PLOS ONE 6, e21896. doi:10.1371/journal.pone.0021896.

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

Domingos, P., 2012. A few useful things to know about machine learning. Communications of the ACM 55, 78. doi:10.1145/2347736.2347755.

Dukart, J., Sambataro, F., Bertolino, A., 2015. Accurate prediction of conversion to Alzheimer's disease using imaging, genetic, and neuropsychological biomarkers. Journal of Alzheimer's Disease 49, 1143–1159. doi:10.3233/JAD-150570. 00022.

Efron, B., 1983. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78, 316–331. doi:10.1080/01621459.1983.10477973.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate Data Analysis and Machine Learning in Alzheimer's Disease with a Focus on Structural Magnetic Resonance Imaging. Journal of Alzheimer's Disease 41, 685–708. doi:10.3233/JAD-131928.

Fleisher, A., Sun, S., Taylor, C., Ward, C., Gamst, A., Petersen, R., Jack, C., Aisen, P., Thal, L., 2008. Volumetric MRI vs clinical predictors of Alzheimer disease in mild cognitive impairment. Neurology 70, 191–199. doi:10.1212/01.wnl.0000287091.57376.65. 00178.

Gavidia-Bovadilla, G., Kanaan-Izquierdo, S., Mataroa-Serrat, M., Perera-Lluna, A., 2017. Early prediction of Alzheimer's disease using null longitudinal model-based classifiers. PLoS ONE 12. doi:10.1371/journal.pone.0168011.

Hall, A., Mattila, J., Koikkalainen, J., Lötjonen, J., Wolz, R., Scheltens, P., Frisoni, G., Tsolaki, M., Nobili, F., Freund-Levi, Y., Minthon, L., Frölich, L., Hampel, H.,

Visser, P., Soininen, H., 2015. Predicting progression from cognitive impairment to alzheimer's disease with the disease state index. Current Alzheimer Research 12, 69–79. doi:10.2174/1567205012666141218123829.

Haller, S., Lovblad, K.O., Giannakopoulos, P., 2011. Principles of Classification Analyses in Mild Cognitive Impairment (MCI) and Alzheimer Disease. Journal of Alzheimer's Disease 26, 389–394. doi:10.3233/JAD-2011-0014.

Iddi, S., Li, D., Aisen, P.S., Rafii, M.S., Thompson, W.K., Donohue, M.C., for the Alzheimer's Disease Neuroimaging Initiative, 2019. Predicting the course of Alzheimer's progression. Brain Informatics 6, 6. URL: https://doi.org/10.1186/s40708-019-0099-0, doi:10.1186/s40708-019-0099-0.

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

Johnson, P., Vandewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P., Macaulay, L., Ellis, K., Szoeke, C., Martins, R., Rowe, C., Masters, C., Ames, D., Zhang, P., 2014. Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease. BMC Bioinformatics 15. doi:10.1186/1471-2105-15-S16-S11. 00027.

Kauppi, K., Fan, C., McEvoy, L., Holland, D., Tan, C., Chen, C.H., Andreassen, O., Desikan, R., Dale, A., 2018. Combining polygenic hazard score with volumetric MRI and cognitive measures improves prediction of progression from mild cognitive impairment to Alzheimer's disease. Frontiers in Neuroscience 12. doi:10.3389/fnins.2018.00260.

Koval, I., Bône, A., Louis, M., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., CHARLIER, B., Bertrand, A., Epelbaum, S., Colliot, O., Allassonnière, S., Durrleman, S., 2018. Simulating Alzheimer's disease progression with personalised digital brain models. URL: https://hal.inria.fr/hal-01964821. preprint.

Leandrou, S., Petroudi, S., Kyriacou, P., Reyes-Aldasoro, C., Pattichis, C., 2018. Quantitative MRI Brain Studies in Mild Cognitive Impairment and Alzheimer's Disease: A Methodological Review. IEEE Reviews in Biomedical Engineering 11, 97–111. doi:10.1109/RBME.2018.2796598.

Li, Y., Yao, Z., Zhang, H., Hu, B., for, t.A.D.N.I., 2018. Indirect relation based individual metabolic network for identification of mild cognitive impairment. Journal of Neuroscience Methods 309, 188–198. doi:10.1016/j.jneumeth.2018.09.007.

Lin, W., Tong, T., Gao, Q., Guo, D., Du, X., Yang, Y., Guo, G., Xiao, M., Du, M., Qu, X., 2018. Convolutional neural networks-based MRI image analysis for the Alzheimer's disease prediction from mild cognitive impairment. Frontiers in Neuroscience 12. doi:10.3389/fnins.2018.00777.

Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S., Alexander, D.C., Consortium, t.E., Initiative, f.t.A.D.N., 2018. TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease. arXiv preprint arXiv:1805.03909 .

Minhas, S., Khanum, A., Riaz, F., Khan, S., Alvi, A., 2018. Predicting progression from mild cognitive impairment to Alzheimer's disease using autoregressive modelling of longitudinal and multimodal biomarkers. IEEE Journal of Biomedical and Health Informatics 22, 818–825. doi:10.1109/JBHI.2017.2703918.

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

Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage 155, 530–548. doi:10.1016/j.neuroimage.2017.03.057.

Samper-Gonzalez, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., Routier, A., Guillon, J., Bacci, M., Wen, J., et al., 2018. Reproducible evaluation of classification methods in alzheimer's disease: Framework and application to mri and pet data. NeuroImage 183, 504–521.

Samper-Gonzalez, J., Burgos, N., Bottani, S., Habert, M.O., Evgeniou, T., Epelbaum, S., Colliot, O., 2019. Reproducible evaluation of methods for predicting progression to Alzheimer's disease from clinical and neuroimaging data, in: Angelini, E.D., Landman, B.A. (Eds.), Medical Imaging 2019: Image Processing, SPIE, San Diego, United States. p. 30. doi:10.1117/12.2512430.

Sarica, A., Cerasa, A., Quattrone, A., 2017. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Frontiers in Aging Neuroscience 9, 329. doi:10.3389/fnagi.2017.00329.

Schuster, C., Elamin, M., Hardiman, O., Bede, P., 2015. Presymptomatic and longitudinal neuroimaging in neurodegeneration—from snapshots to motion picture: a systematic review. Journal of Neurology, Neurosurgery & Psychiatry 86, 1089–1096. doi:10.1136/jnnp-2014-309888.

Thung, K.H., Yap, P.T., Adeli, E., Lee, S.W., Shen, D., 2018. Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. Medical Image Analysis 45, 68–82. doi:10.1016/j.media.2018.01.002.

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., 2017. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease. IEEE Transactions on Biomedical Engineering 64, 155–165. doi:10.1109/TBME.2016.2549363.

Wen, J., Thibeau-Sutre, E., Samper-González, J., Routier, A., Dormont, D., Durrleman, S., Colliot, O., Burgos, N., 2019. How serious is data leakage in deep learning studies on Alzheimer's disease classification?, in: Proceedings of: Human Brain Mapping (HBM), p. 8.

Wooller, S., 2018. Artificial intelligence can predict Alzheimer's 6 years earlier than medics, study finds. URL: https://www.foxnews.com/health/artificial-intelligence-can-predict-alzheimers-6-years-earlier-than-medics-study-finds.

Appendix A. Supplementary Materials

Appendix A.1. Query

The full query was:

```
TITLE-ABS-KEY ("alzheimer's" OR alzheimer OR ad) AND TITLE-ABS-
    KEY ("Mild Cognitive Impairment" OR "MCI") AND TITLE-ABS-
    KEY ((predicting OR prediction OR predictive) AND (
    conversion OR decline OR progression OR onset) OR prognosis
    ) AND TITLE-ABS-KEY (accuracy OR roc OR auc OR specificity
    OR sensitivity) AND (TITLE-ABS-KEY ("Deep learning" OR "
    neural network" OR "neural networks" OR "convolutional
    network" OR "convolutional networks" OR "bayesian network"
    OR "bayesian networks") OR TITLE-ABS-KEY ("Matrix
    completion" OR "Support vector machine" OR "linear mixed-
    effect" OR "logistic regression" OR "Random Forest" OR "
    kernel classifier" OR "kernel" OR "decision tree" OR "
    decision trees" OR "least-squares") OR TITLE-ABS-KEY ("
    Machine learning" OR "pattern recognition" OR "pattern
    classification" OR "classifier" OR "algorithm" OR "
    classification"))
```

Appendix A.2. Selection process diagram

The process used to select the articles included in the review is shown in Figure S1.

Appendix A.3. Reported items

For each article, the following elements were reported:

- number of MCI subjects progressing to AD;
- number of stable MCI subjects;
- time to prediction;
- used cohorts;
- use of socio-demographic features (yes/no);
- use of APOE (yes/no);
- use of general cognitive features (yes/no);
- use of domain-targeted cognitive features (yes/no);
- use of new, home-made cognitive features (yes/no);
- use of voxel based features from T1 MRI (yes/no);

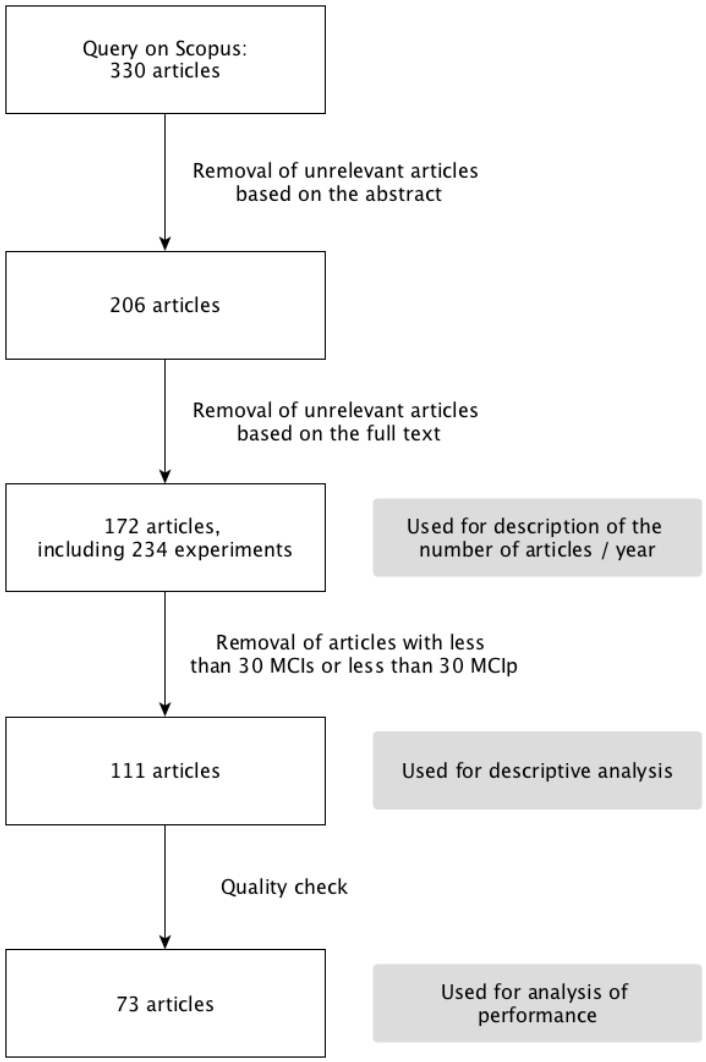


Figure S1: Diagram representing who the articles were selected

- use of regions of interest on the whole brain, from T1 MRI (yes/no);
- use of selected regions of interest from T1 MRI (yes/no);
- use of white matter hyper-intensities (yes/no);
- use of PET FDG features (yes/no);
- use of PET amyloid features (yes/no);
- use of PET tau features (yes/no);
- use of CSF features (yes/no);
- use of amyloid status (yes/no);
- use of DTI features (yes/no);
- use of functional MRI features (yes/no);
- use of EEG or MEG features (yes/no);
- use of other features (yes/no, precision given as a free note);
- use of longitudinal features (yes/no);
- is feature selection performed (yes/no);
- used algorithm (categories defined below);
- validation method (categories defined bellow);
- feature selection performed on the whole data set (yes/no/unclear);
- feature embedding performed on the whole data set (yes/no/unclear);
- selection of the input visit of the test subjects using their date of progression to AD (yes/no);
- other data leakage (use of the test set to make decisions) (yes/no/unclear);
- other issue (yes/no)
- AUC value;
- accuracy value;
- balanced accuracy value;
- sensitivity value;
- specificity value;

Free notes describing the issues, or important points that did not fit in the previous list, were added.

The possible algorithm categories were added by the readers and aggregated. The final list was: bayesian algorithms, classification by clinicians, gaussian process, linear discriminant analysis (LDA), low rank matrix completion (LRMC), linear regression, logistic regression, manifold learning, multiple kernel learning, neural network, orthogonal partial least square (OPLS), random forest, regularized logistic regression, support vector machine, survival analysis, use of a threshold and others (including home-made algorithms).

The same process was used to create the cross-validation category list, composed of: 10-fold, k-fold, repeated k-fold, leave one out, out of the bag, single split, repeated single split, validation on an independent cohort, validation on different groups (when the algorithm is trained on separating AD and CN subjects, and tested on predicting the progression of MCI subjects), none, not described (when the use of cross-validation is mentioned but the used validation method is not described) and not needed (for thresholding with a manually chosen threshold for example).

Appendix A.4. Journals and conference proceedings

Table S1 shows the journals and conference proceedings in which more than one included article has been published, and the associated number of articles.

Journal or conference proceedings	**Number of included articles**
Journal of Alzheimer's Disease	12
NeuroImage	11
Lecture Notes in Computer Science	7
PLoS ONE	9
Neurobiology of Aging	6
Neurology	3
Brain Topography	3
Current Alzheimer Research	3
Medical Image Analysis	3
Frontiers in Aging Neuroscience	3
Scientific Reports	2
Frontiers in Neuroscience	2
IEEE Journal of Biomedical and Health Informatics	2
IEEE Transactions on Biomedical Engineering	2
NeuroImage: Clinical	2
Journal of Neuroscience Methods	2

Table S1: Number of included articles published in each journal or conference proceedings. Only the journals with more than one included article are shown here. The articles taken into account are the one considered for analysis, and that use a large enough data set.

Appendix A.5. Information table

A table containing all the articles included in the review and all the reported values can be found on `https://gitlab.com/icm-institute/aramislab/mci-progression-review`. The issues identified in each articles were removed from this open-access table, to avoid negatively pointing at these studies. They can be made available if requested to the corresponding author.