Prediction of amyloidosis from neuropsychological and MRI data for cost effective inclusion of pre-symptomatic subjects in clinical trials

Manon Ansart, Stéphane Epelbaum, Geoffroy Gagliardi, Olivier Colliot, Didier Dormont, Bruno Dubois, Harald Hampel, Stanley Durrleman

Prediction of amyloidosis from neuropsychological and MRI data for cost effective inclusion of pre-symptomatic subjects in clinical trials

Manon Ansart[1,2], Stéphane Epelbaum[1,2,3], Geoffroy Gagliardi[1,3], Olivier Colliot[1,2,3,4], Didier Dormont[1,2,4], Bruno Dubois[1,3], Harald Hampel[1,3,5], Stanley Durrleman[2,1], for the ADNI, and the INSIGHT study group

[1] Sorbonne Universités, UPMC Univ Paris 06, Inserm, CNRS, Institut du cerveau et de la moelle (ICM) - Pitié-Salpêtrière hospital, Bvd de l'hôpital, Paris, France
[2] Inria Paris, Aramis project-team, 75013, Paris, France
[3] AP-HP, Pitié-Salpêtrière hospital, Department of Neurology, Institut de la Mémoire et de la Maladie d'Alzheimer (IM2A), Paris, France
[4] AP-HP, Pitié-Salpêtrière hospital, Department of Neuroradiology, Paris, France
[5] AXA Research Fund & UPMC Chair, Paris, France

Abstract. We propose a method for selecting pre-symptomatic subjects likely to have amyloid plaques in the brain, based on the automatic analysis of neuropsychological and MRI data and using a cross-validated binary classifier. By avoiding systematic PET scan for selecting subjects, it reduces the cost of forming cohorts of subjects with amyloid plaques for clinical trials, by scanning fewer subjects but increasing the number of recruitments. We validate our method on three cohorts of subjects at different disease stages, and compare the performance of six classifiers, showing that the random forest yields good results more consistently, and that the method generalizes well when tested on an unseen data set.

1 Introduction

One of the lesions defining Alzheimer's disease (AD) is the formation of amyloid plaques in the brain. A commonly accepted hypothesis is that this plaque formation is the starting point that triggers a cascade of events leading to neuronal loss, cognitive decline and then dementia[9]. Those plaques appear very early in the disease course, often way before any signs of cognitive decline and diagnosis [5, 12]. They are the consequence of the aggregation of beta-amyloid (Aβ) peptides, and together with neurofibrillary tangles, they are thought to cause the death of neurons, hence being the potential cause of cognitive decline.

Consequently, amyloid plaques are targeted by several molecules at different stages of their formation, with the aim that preventing their formation or clearing them would stop the process resulting in AD. Several of those potential drugs, such as solanezumab and bapineuzumab have already been tested on mild-to-moderate AD subjects and did not prove to slow down the progression of the symptoms of AD [4]. A possible explanation for these failures is that the

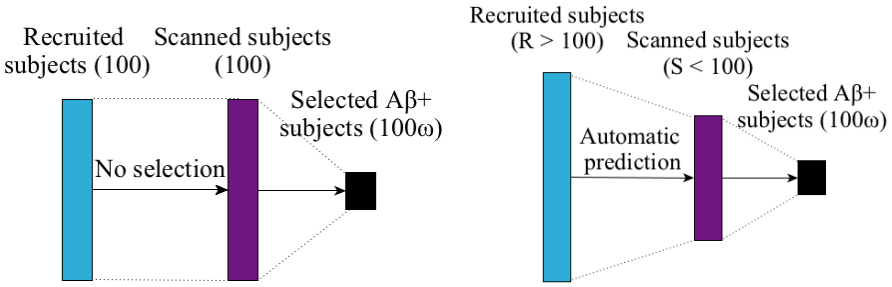

Fig. 1: Current (left) and proposed (right) processes for Aβ+ subjects selection

treatments have been tested on subjects too late in the disease course, and for some of the trials on subjects without confirmation of amyloidosis. Drugs may stop the formation of amyloid plaques or clear them effectively, but they cannot repair the damage that has already been caused by the plaques. An hypothesis is that applying those treatments specifically on pre-symptomatic subjects with amyloidosis would make them more effective.

Testing these molecules at the preclinical stage raises the problem of recruiting pre-symptomatic subjects with amyloid plaques [15]. Positron emission tomography (PET) imaging with amyloid ligands is, together with lumbar puncture, the most widely used techniques to assess amyloid plaques presence *in vivo*. However, the prevalence of Aβ positive (Aβ+) subjects among asymptomatic elderly people is rather low: about 30 % [2], resulting in 333 subjects to recruit and scan to get 100 Aβ+ subjects. A PET scan is however quite costly, about 1,000€ in Europe, and 5,000$ in the USA, so creating cohorts of pre-symptomatic Aβ+ subjects amounts to be very expensive. To ease the economic burden, we propose here to introduce a pre-screening phase to select subjects with higher risk of being Aβ+ than in the general population, and perform a confirmatory PET scan to those subjects only. We propose to predict the presence of amyloidosis in subjects by the automatic analysis of their neuropsychological assessments and structural imaging data, which are exams that are less expensive. We propose to use machine learning algorithms to find the patterns in these data that best predict the presence of amyloid plaques in the brain.

Methods to automatically predict amyloidosis from socio-demographic, genetic and cognitive variables have been proposed in [13] and [11]. In particular, they studied how univariate methods perform compared with multivariate ones. The threshold of the logistic regression in [13] was set a priori, and not optimized. The approach in [11] aimed to maximize the Positive Predictive Value (PPV). This value might be arbitrarily good by using a more and more stringent detection threshold, but that implies that more and more subjects need to be recruited for a given target number of Aβ+ subjects, as many positive subjects are discarded as false negatives. To better reflect this trade-off, we propose to translate the specificity and sensitivity of a classifier into a number of subjects to be recruited (R) and a number of subjects to be scanned (S), as shown Fig. 1.

Each value of R and S corresponds to a given cost for the constitution of the cohort. The aim of our approach is to find the threshold minimizing this cost.

In this paper, we will benchmark an array of cross-validated machine learning algorithms for the prediction of amyloidosis from several feature sets extracted from clinical and structural imaging data. We will validate these algorithms on three different cohorts with subjects at various disease stages. The hyper-parameters will be tuned by maximizing the area under the ROC curve (AUC). The score threshold will be chosen so as to minimize the cost.

2 Materials and Methods

2.1 Validation cohorts

The method is validated on 3 cohorts: INSIGHT, ADNI-CN and ADNI-MCI. INSIGHT is a monocentric French study including asymptomatic subjects with a subjective memory complaint (SMC). 318 subjects have an AV45 PET scan and hence an Aβ standardized uptake value ratio (SUVr) for baseline, among which 88 (27.7%) are Aβ+.

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multicentric longitudinal study. We use the cognitively normal subjects (ADNI-CN) and the subjects with mild cognitive impairments (ADNI-CN) that have an Aβ status assessed by AV45 PET scan or CSF biomarkers in the absence of PET scan. The baseline visit of the subjects who stay Aβ+ or Aβ- for all visits is used. 431 CN subjects (37.6% of Aβ+) and 596 MCI subjects (62.9% of Aβ+) are available.

2.2 Input Features

Socio-demographic (age, gender, education), genetic (APOE) and cognitive features are used as inputs. For ADNI, the Alzheimer's Disease Assessment Scale cognitive sub-scale (ADAScog) is divided into memory, language, concentration and praxis, and for INSIGHT SMC questionnaires and cognitive tests (targeting memory, executive functions, behavior or overall cognitive skills) are used. MRI features are also used and compared with cognitive assessments in terms of prediction power. Cortical thicknesses averaged on 72 regions are extracted using FreeSurfer and divided by the total cortical thickness. The hippocampal volume is computed using FreeSurfer for ADNI and SACHA [3] for INSIGHT.

2.3 Algorithms

The classification is made using different algorithms in order to compare their performance. Hyper-parameters are tuned using cross-validation to maximize the AUC. The used algorithms are: random forest [1] (validation of the number and depth of the trees), regularized logistic regression [8] (validation of the regularization parameter), linear support vector machine [14] (SVM) (validation of the penalty parameter), additive logistic regression [7] (AdaLogReg) (validation

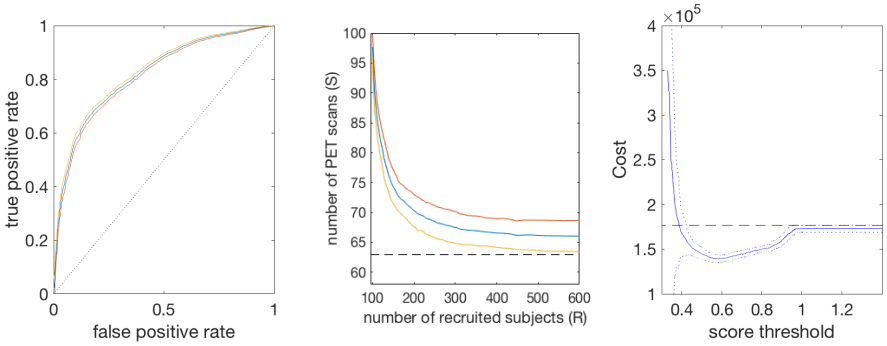

Fig. 2: Example of ROC curve (left), S vs R curve (middle) and corresponding cost curve (right)

of the number and depth of the learners and of the learning rate for shrinkage), and adaptive boosting [6] (AdaBoost) (same hyper-parameters as AdaLogReg).

The data set is randomly split into a training (70%) and a test set (30%) 50 times, and a 5-fold validation is performed on the training set to automatically choose the algorithm hyper-parameters. All algorithms are trained on the whole training set, and their performance is evaluated on the test set. The performance mean and standard deviation (std) are computed and used to perform t-tests.

2.4 Performance Measures

The AUC is used to evaluate the overall performance of the methods and to tune the hyper-parameters. The maximum balanced accuracy (average of sensitivity and specificity, noted BAcc), which corresponds to a specific point on the ROC curve, is also used. The last measure is the minimal cost for recruiting $100 * \omega$ Aβ+ subjects, where ω is the proportion of positive subjects in the data set. In order to compute this cost, the ROC curve is computed (Fig 2, left), then the number of subjects that have to be recruited (R) and the number of subjects that have to be scanned (S) are computed for each point on the ROC curve (Fig 2, middle):

$$S = 100 * \omega * \frac{\mathrm{TP} + \mathrm{FP}}{\mathrm{TP}} \qquad (1) \qquad\qquad R = 100 * \omega * \frac{\mathrm{N}}{\mathrm{TP}} \qquad (2)$$

where TP stands for number of True Positives, FP for number of False Positives and N for number of tested subjects. The corresponding cost is computed at each point (Fig 2, right). The point with the minimal cost is kept, and the corresponding cost is used as a performance measure. We made the hypothesis that recruiting a subject (with cognitive scores and genetic information) costs 100€, doing an MRI costs 400€ and a PET scan 1,000€. As a comparison, recruiting $100 * \omega$ Aβ+ subjects doing a confirmatory PET scan for all subjects would correspond to a cost of 110,000€ (100 recruitments and PET scans).

Table 1: Benchmark of algorithms, given in the form: average performance (std)

	Data set	Random Forest	Logistic regression	SVM	AdaLogReg	AdaBoost
AUC	INSIGHT	67.5 (5.5)	62.7 (6.1)	62.0 (5.8)	67.5 (5.7)	67.2 (6.9)
	ADNI-CN	69.1 (4.0)	69.5 (4.1)	67.3 (5.0)	66.4 (4.6)	66.5 (5.1)
	ADNI-MCI	83.8 (2.8)	82.5 (2.6)	82.4 (2.7)	82.6 (2.8)	83.1 (3.3)
BAcc	INSIGHT	63.9 (1.5)	60.1 (1.6)	59.6 (1.4)	62.3 (1.5)	62.3 (1.3)
	ADNI-CN	63.3 (1.0)	63.8 (1.1)	62.3 (0.9)	61.6 (1.3)	61.4 (1.3)
	ADNI-MCI	74.5 (0.9)	74.2 (0.8)	74.5 (0.8)	73.6 (0.8)	73.4 (1.0)
Cost (€)	INSIGHT	80,697 (15,900)	91,866 (15,811)	96,813 (13,147)	85,134 (16,137)	85,118 (18,944)
	ADNI-CN	88,206 (86,88)	84,833 (7,694)	88,404 (9,049)	93,231 (8,216)	92,921 (9,192)
	ADNI-MCI	85,673 (30,56)	86,460 (2,522)	86,436 (2,642)	86,056 (2,566)	86,269 (3,485)

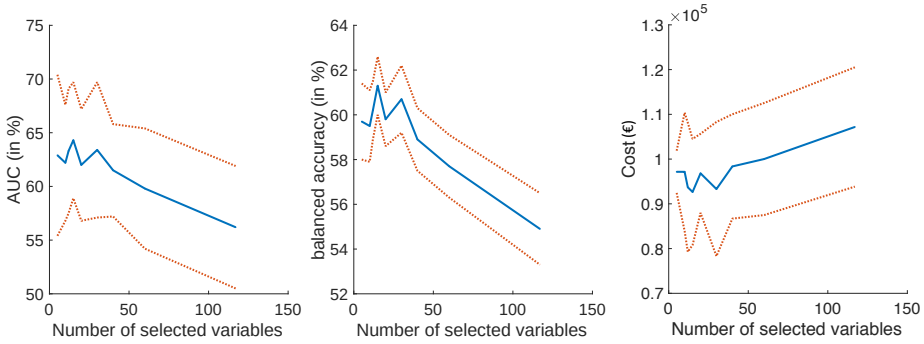

Fig. 3: Performance variations depending on the number of kept lasso variables

3 Experiments and Results

3.1 Algorithm benchmark

An algorithm benchmark (using socio-demographic, genetic and cognitive features) is presented in Table 1. There is no algorithm that consistently outperforms the others for all criterion. However, if a choice has to be made, the random forest is consistently among the best algorithms for all measures and all data sets. Its performances are the best for INSIGHT and ADNI-MCI, and are slightly below the ones of the logistic regression for the ADNI-CN. Using a random forest leads to a significant decrease in the cost of recruiting $100 * \omega$ Aβ+ compared to the initial cost of 110,000€ ($p<0.001$).

3.2 Feature selection

Using all the INSIGHT available features (117 features including 112 cognitive ones) for prediction gives an AUC of 56.2% ($\pm$7.5). Dimension reduction is

Table 2: Results using MRI variables, socio-demographic and genetic information on different data sets

Data set	AUC in % (std)	BAcc in % (std)	cost in € (std)
Trained and tested on INSIGHT	61.9 (6.5)	59.3 (1.5)	14,6147 (4,975)
Trained on ADNI-CN, tested on INSIGHT	62 (6.6)	58 (1.7)	14,5989 (5,112)
Trained on ADNI-CN, tested on INSIGHT (all samples)	66.1 (3.6)	62.5 (1.1)	14,5896 (2,663)
Trained and tested on [INSIGHT ADNI-CN]	61.3 (6.6)	58.5 (1.5)	14,5642 (3,897)
Trained and tested on [INSIGHT ADNI-CN] (all samples)	66.7 (3.7)	62.3 (1.0)	14,6613 (5,859)

therefore considered, comparing several methods. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) using fastICA[10] are first considered with a variable number of selected dimensions, but both give less than 52% of AUC. Alternatively, Lasso feature selection is performed, using a linear regression followed by a random forest and keeping from 5 up to 60 features (Fig. 3). The best results, obtained on 15 features, correspond to an AUC of 64.3% ($\pm$5.2), which is significantly better than using all features ($p<0.001$). Another strategy is forming aggregates for each cognitive test using expert knowledge on which test variables are most likely to be a marker of AD. 26 cognitive summary variables are constructed this way, and using them in place of the 112 original cognitive features gives an AUC of 67.5% ($\pm$5.5), which is significantly better ($p<0.005$) than the performance reached using automatic methods.

3.3 Use of MRI

Using socio-demographic, genetic and cognitive features yields an AUC of 67.5% ($\pm$5.5) on INSIGHT (Table 1, column 1). Using MRI instead of cognitive features leads to a decrease in AUC (Table 2 line 1: 61.9 % $\pm$6.5, $p<0.001$). Using both results in a non-significant increase in AUC (68.8 $\pm$4.4, $p>0.1$), and in a cost increase, as it implies to do an MRI on all potential subjects. The performance vs cost ratio is therefore better without MRI.

3.4 Generalization on an independent cohort

INSIGHT and ADNI are diferent databases, as INSIGHT is a monocentric study focused on SMC, and ADNI is multicentric with different inclusion criterion and goals. The hippocampal volumes have also been extracted using different softwares. In order to see if the proposed method could generalize well to other data sets, it is trained on ADNI-CN and tested on INSIGHT, as they correspond to the most similar subject profiles. The socio-demographic, genetic and MRI variables are used, and a lasso selection of 12 features is performed on the MRI

variables. In order to have a fair comparison, training and test set are created with the same size as the training and test data sets coming from INSIGHT, by randomly selecting $318 * 0.7 = 223$ subjects from ADNI-CN for the training set and $318 * 0.3 = 95$ from INSIGHT for the test set. This sampling and the classification are performed 50 times in order to get an average performance. The performances obtained by learning on either INSIGHT alone (Table 2 line 1) or ADNI-CN (Table 2 line 2) are very similar, which means the proposed method is likely to give similar results if applied on a new data set of CN elderly subjects.

3.5 Pooling data sets

A new data set is created by pooling subjects from ADNI-CN and INSIGHT, while keeping the same total cohort size as in INSIGHT. The method gives similar performances when validated on this pooled data set or on INSIGHT (Table 2, lines 1 and 4), which shows that the heterogeneity of pooled data sets does not alter the classification performances.

3.6 Effect of sample size

When the classifier is trained and tested on INSIGHT, $318*0.7 = 223$ subjects are used for training. The training set can contain up to 431 subjects when training on ADNI and testing on INSIGHT, and 524 subjects when using the pooled data set, which is respectively 2.30 and 1.86 times larger. We can therefore train the method on larger and larger data sets, keeping the same proportion between the training and the test set (70%-30%) for comparison. The results, reported in Table 2 lines 3 and 5, show a significant increase in the AUC ($p<0.001$) when the size of the data set increases, which comforts the need to create large databases, or pool existing databases, to create more accurate medical models.

4 Conclusion

We proposed a method for creating cohorts of Aβ+ pre-symptomatic subjects, by building a classifier optimized to minimize cohort creation costs. The proposed method identifies in a pre-screening phase a sub-set of subjects with a much higher prevalence of Aβ+ cases. We benchmarked cross-validated algorithms and showed that the random forest consistently yields good results. We tested our method on 3 data sets and showed that it always results in a significant cost decrease for creating such cohorts. We showed that the method generalizes well when trained on a cohort and tested on an independent one, therefore showing its potential for being used in real clinical environment with heterogeneous procedures for subject selection, data acquisition and processing. The best costs are achieved by using socio-demographic, genetic and cognitive features chosen using expert knowledge. Using MRI features increases the overall costs, but the performances could be increased by extracting more complex features, or by using a priori knowledge for selecting relevant variables.

This work was partly funded by ERC grant N°678304, H2020 EU grant N°666992 and ANR grant ANR-10-IAIHU-06. HH is supported by the AXA Research Fund, the Fondation UPMC and the Fondation pour la Recherche sur Alzheimer, Paris, France. OC is supported by a "contrat d'interface local" from AP-HP.

References

1. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
2. G. Chetelat, R. La Joie, N. Villain, A. Perrotin, V. de La Sayette, F. Eustache, and R. Vandenberghe. Amyloid imaging in cognitively normal individuals, at-risk populations and preclinical alzheimer's disease. *Neuroimage Clin*, 2:356–365, 2013.
3. M. Chupin, A. Hammers, R. S. N. Liu, O. Colliot, J. Burdett, E. Bardinet, J. S. Duncan, L. Garnero, and L. Lemieux. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: method and validation. *NeuroImage*, 46(3):749–761, 2009.
4. R. S. Doody, R. G. Thomas, M. Farlow, T. Iwatsubo, B. Vellas, S. Joffe, K. Kieburtz, R. Raman, X. Sun, P. S. Aisen, E. Siemers, H. Liu-Seifert, and R. Mohs. Phase 3 trials of solanezumab for mild-to-moderate alzheimer's disease. *New England Journal of Medicine*, 370(4):311–321, 2014.
5. B. Dubois, H. Hampel, H. H. Feldman, P. Scheltens, P. Aisen, S. Andrieu, H. Bakardjian, H. Benali, L. Bertram, K. Blennow, K. Broich, E. Cavedo, S. Crutch, J.F. Dartigues, C. Duyckaerts, S. Epelbaum, G. B. Frisoni, S. Gauthier, R. Genthon, and A. A. Gouw et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. *Alzheimer's & Dementia*, 12(3):292–323, 2016.
6. J. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.
7. J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *The annals of statistics*, 28(2):337–407, 2000.
8. J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.
9. J. A. Hardy and G. A. Higgins. Alzheimer's disease: the amyloid cascade hypothesis. *Science*, 256(5054):184–185, 1992.
10. A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.
11. P. S. Insel, S. Palmqvist, R. S. Mackin, R. L. Nosheny, O. Hansson, M. W. Weiner, and N. Mattsson. Assessing risk for preclinical β-amyloid pathology with APOE, cognitive, and demographic information. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 4:76–84, 2016.
12. C. R Jack, D. S Knopman, W. J Jagust, L. M Shaw, P. S Aisen, M. W Weiner, R. C Petersen, and J. Q Trojanowski. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *Lancet Neurology*, 9(1):119, 2010.
13. M. M. Mielke, H. J. Wiste, S. D. Weigand, D. S. Knopman, V. J. Lowe, R. O. Roberts, Y. E. Geda, Dana M. Swenson-Dravis, B. F. Boeve, M. L. Senjem, P. Vemuri, R. C. Petersen, and C. R. Jack. Indicators of amyloid burden in a population-based study of cognitively normal elderly. *Neurology*, 79(15):1570–1577, 2012.
14. K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.
15. J. T. O'Brien and K. Herholz. Amyloid imaging for dementia in clinical practice. *BMC Medicine*, 13, 2015.