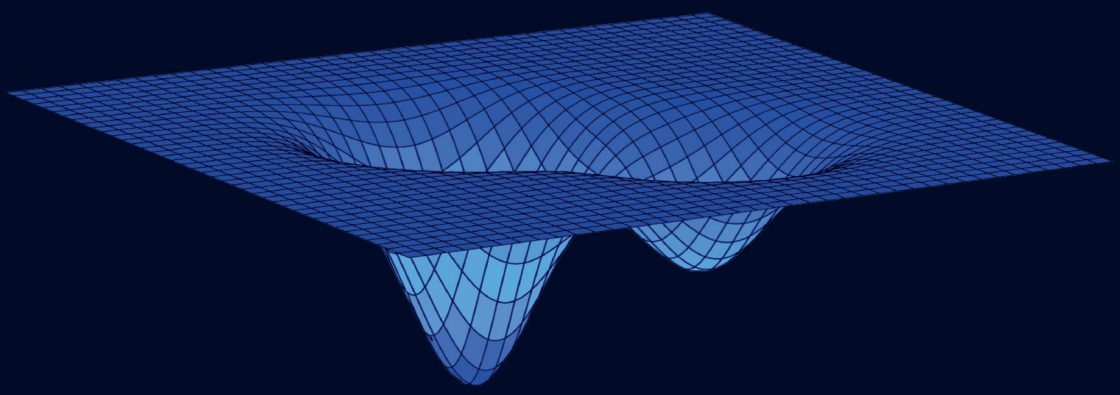
Supervised machine learning in psychiatry

SUPERVISED MACHINE LEARNING IN PSYCHIATRY

TOWARDS APPLICATION IN CLINICAL PRACTICE



MASSIMILIANO GRASSI

SUPERVISED MACHINE LEARNING IN PSYCHIATRY: TOWARDS APPLICATION IN CLINICAL PRACTICE

By
Massimiliano Grassi

The work presented in this thesis was performed at the Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University, Maastricht, the Netherlands.

ISBN: 978-94-6423-915-7
DOI: 10.26481/dis.20220919mg

Cover design: Laura Poluzzi
Printing: Proefschriftmaken, The Netherlands

SUPERVISED MACHINE LEARNING IN PSYCHIATRY: TOWARDS APPLICATION IN CLINICAL PRACTICE

Dissertation:
to obtain the degree of
Doctor at the Maastricht University,
on the authority of the Rector Magnificus,
Prof.dr. Pamela Habibović
in accordance with the decision of the Board of Deans
To be defended in public on
Monday 19th of September 2022, at 16:00 hours

By
Massimiliano Grassi

Supervisors:

- prof.dr. K.R.J. Schruers
- prof.dr. M. Dumontier

Co-supervisor:

- prof.dr. G. Perna, Professor of Psychiatry, Humanitas University, Rozzano, Milan, Italy

Assessment Committee:

- prof.dr. T.A.M.J. van Amelsvoort (Chair)
- dr. A. Coulet, French National Research Institute for Computer Science (INRIA), Le Chesnay, France
- prof.dr. P.A.E.G. Delespaul
- prof.dr. G.J. Hendriks, Radboud University Medical Center Nijmegen

Perhaps solving the riddle of the universe requires
one more neuron than, de facto,
anyone will ever have.

Jerry A. Fodor

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION[1]

1 Part of the paragraphs 1.1 and 1.2 has been extracted and adapted from an editorial by Perna, Grassi, and other Authors Perna, G., M. Grassi, D. Caldirola and C. Nemeroff (2018). "The revolution of personalized psychiatry: will technology make it happen sooner?" Psychological medicine **48**(5): 705-713.

CHAPTER 2

A CLINICALLY TRANSLATABLE MACHINE LEARNING ALGORITHM FOR THE PREDICTION OF AD CONVERSION IN INDIVIDUALS WITH MILD AND PREMILD COGNITIVE IMPAIRMENT

Massimiliano Grassi[1], Giampaolo Perna[1,2,3,4], Daniela Caldirola[1], Koen Schruers[2], Ranjan Duara[3,5,6], David A. Loewenstein[3,6,7]

1 Department of Clinical Neurosciences, Hermanas Hospitalarias - Villa San Benedetto Menni Hospital, Italy.
2 Research Institute of Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine and Life Sciences, University of Maastricht, Maastricht, Netherlands.
3 Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, FL, USA.
4 Mantovani Foundation, Arconate, Italy.
5 Department of Neurology, Herbert Wertheim College of Medicine, Florida International University, Miami, FL, USA.
6 Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami, FL, USA.
7 Center on Aging, Miller School of Medicine, University of Miami, FL, USA.

Reference: Grassi M, Perna G, Caldirola D, Schruers K, Duara R, Loewenstein DA. A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. J Alzheimers Dis. 2018;61(4):1555-1573.

Abstract

Background: available therapies for Alzheimer's disease can only alleviate and delay the advance of symptoms, with the greatest impact eventually achieved when provided at an early stage. Thus, early identification of which subjects at high risk, e.g., with MCI, will later develop Alzheimer's disease is of key importance. Currently available machine learning algorithms achieve only limited predictive accuracy or they are based on expensive and hard-to-collect information.
Objective: the current study aims to develop an algorithm for a 3-years prediction of conversion to Alzheimer's disease in MCI and PreMCI subjects based only on non-invasively and effectively assessable predictors.
Methods: a dataset of 123 MCI/PreMCI subjects was used to train different machine learning techniques. Baseline information regarding sociodemographic characteristics, clinical and neuropsychological test scores, cardiovascular risk indexes, and a visual rating scale for brain atrophy was used to extract 36 predictors. Leave-pair-out-cross-validation was employed as validation strategy and a recursive feature elimination procedure was applied to identify a relevant subset of predictors.
Results: 16 predictors were selected from all domains excluding sociodemographic information. The best model resulted a support vector machine with radial-basis function kernel (whole sample: AUROC = 0.962, best balanced accuracy = 0.913; MCI sub-group alone: AUROC = 0.914, best balanced accuracy = 0.874).
Conclusions: Our algorithm shows very high cross-validated performances that outperform the vast majority of the currently available algorithms although only non-invasive and effectively assessable predictors are used. Further testing and optimization in independent samples will warrant its application in both clinical practice and clinical trials.

Keywords: Alzheimer's disease, clinical prediction rule, machine learning, mild cognitive impairment, personalized medicine.

Introduction

Alzheimer's disease is a neurodegenerative disease characterized by progressive loss of memory and functional abilities that leads to severe dementia and eventually death. It is the most common neurodegenerative disease and currently affects 47 million people worldwide, being the top cause for disabilities in later life. The global cost of Alzheimer's and dementia is estimated to be $818 billion, which is nearly the 1% of the entire world's gross domestic product. These numbers are projected to increase, with a global expected cost of $2 trillion by 2030 and more than 131 million people suffering from this disorder by 2050 (International 2016).

No cure or disease modifying treatment is currently available for Alzheimer's disease and current treatment regimens only provide symptomatic relief (Szeto and Lewis 2016). By the time Alzheimer's disease is clinically diagnosed, there is considerable multi-system degeneration that has occurred within the brain. As such, emerging treatments will likely have the greatest impact when provided at an earliest possible stages of the disease process (Brooks and Loewenstein 2010, Loewenstein, Curiel et al. 2017).

Therefore, the prompt identification of subjects truly at high risk of developing Alzheimer's disease is a crucial issue still without a solution.

Mild Cognitive Impairment (MCI) is a condition characterized by changes in cognitive capabilities beyond what is expected for the subject's age and education that are sufficiently mild that they do not interfere significantly with its daily activities. Individuals with such condition are at high risk of converting to dementia and especially Alzheimer's disease in the next few years (20-40% of conversion rate by three years, with a lower rate evidenced in epidemiologic samples than in clinical ones (Petersen, Parisi et al. 2006, Roberts, Knopman et al. 2014).

Furthermore, even subjects with an intermediate state between normal cognition and MCI, i.e., the so called Premild Cognitive Impairment (PreMCI) stage (Chao, Mueller et al. 2010), are more likely to progress to a formal diagnosis of MCI or dementia within a two- to three-year period and this might represent the earliest clinically definable stage of Alzheimer's disease. (Loewenstein, Greig et al. 2012).

However, some subjects with MCI have shown to remain stable over years or even to recover to cognitively normal with no further progression to Alzheimer's disease. This holds even more true for subjects with PreMCI than for those with MCI (Loewenstein, Greig et al. 2012). Different health problems other than neurodegenerative diseases can cause transient MCI and PreMCI conditions and these do not necessarily lead to Alzheimer's disease (Breitner 2014). Thus, sole reliance on these precursor conditions is not enough to provide a precise identification of those subjects at true risk of later developing Alzheimer's disease.

Beyond MCI and preMCI, several attempts to identify subject's characteristics that may improve the prediction of progression to Alzheimer's disease have been done. Investigations have regarded a vast variety of potential predictors, such as sociodemographic and clinical characteristics, cognitive performances, neuropsychiatric symptomatology, cardiovascular indexes, dietary and life habits, structural and functional neuroimaging investigations, gene typization, and several bio-markers assessed both in the cerebrospinal fluid and peripherally (van Rossum, Vos et al. 2010, Klunk 2011, Forlenza, Diniz et al. 2013, Kang, Korecka et al. 2013, Sperling and Johnson 2013, Cooper, Sommerlad et al. 2015, Van Cauwenberghe, Van Broeckhoven et al. 2016).

It is increasingly recognized that better predictive capability can be achieved by models that simultaneously exploit the information coming from several predictors, and machine learning can be used to create such models. This is a fast-growing field at the crossroads of computer science, engineering, and statistics "that gives computers the ability to learn without being explicitly programmed" (Samuel 1959). Machine learning techniques use known training examples to create algorithms able to provide the best possible prediction when applied to new cases whose outcome is still unknown. Machine learning has been applied in the attempt to predict MCI-Alzheimer's disease conversion in more than 50 published studies. Different combinations of the above-mentioned predictors were applied to various machine learning techniques in the attempt to predict conversion from MCI to dementia from one year to even five years after the baseline assessment. The results achieved vary broadly among studies, ranging from some that achieved performances just above the chance to a few showing high accuracy levels (Plant,

Teipel et al. 2010, Apostolova, Hwang et al. 2014, Clark, Kapur et al. 2014, Agarwal, Ghanty et al. 2015, Moradi, Pepe et al. 2015, Dukart, Sambataro et al. 2016, Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017, Minhas, Khanum et al. 2017).

Despite this huge research effort, no gold-standard algorithm is available to predict progression in those at risk for Alzheimer's disease and clinical translation is still lacking. All the "top performing" algorithms have not been tested in further independent samples thus far, and, in addition, certain predictors employed by some models may represent a significant barrier in their clinical adoption due to their high costs and/or invasiveness (e.g., fludeoxyglucose positron emission tomography scans or lumbar puncture).

Considering all the above-mentioned issues, the current study aims to be the first step in the development of a clinically-translatable algorithm for the identification of the Alzheimer's disease conversion in subjects with either MCI or PreMCI. To be quickly adoptable in clinical practice, the algorithm should include only non-invasive predictors that are either already routinely assessed or effectively introducible in clinical practice and achieve a high predictive accuracy. Considering the evidence available so far, we hypothesize that the information provided by sociodemographic characteristics, clinical and neuropsychological tests, cardiovascular risk indexes, and clinician-rated level of brain atrophy might allow for such predictive models. In this investigation, a series of machine learning algorithm will be developed and cross-validated within a sample of patients with either MCI/PreMCI whose diagnostic follow-up was available for at least three years after the baseline assessment. Out-of-the-sample testing of the best algorithm in independent samples of MCI/PreMCI patients will be performed in a further phase.

Materials and methods

Subjects

Data regarding 184 subjects with MCI or PreMCI at baseline and with available diagnostic follow-up assessments for at least three years were included in the study.

These are part of a dataset that collects several patients recruited in a study investigating longitudinal changes associated with MCI and normal aging that involved community volunteers as well as from the Memory Disorders Clinic at the Wien Center for Alzheimer's disease and Memory Disorders at Mount Sinai Medical Center, Miami, Beach, Florida as well as subjects recruited from the community and memory disorders center at the University of South, Florida. All subjects at each of the sites had a common clinical and neuropsychological battery as described below.
Considering the final aim of developing a predictive algorithm to be used in clinical practice, no other inclusion or exclusion criteria were applied beyond these diagnostic criteria.
Subjects were classified as converters to probable Alzheimer's disease (cAD; n = 48, 26,1%) if during at least one of the follow-up assessments occurred within three years from the baseline investigation, they presented a Dementia syndrome by DSM-IV-TR criteria (American Psychiatric Association 2000), and satisfied the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association criteria for Alzheimer's disease (McKhann, Drachman et al. 1984). Otherwise, they were classified as non-converters to Alzheimer's disease (NC; n = 136, 73,9%).
The study was conducted with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All subjects gave their written informed consent to the use of their clinical data for scientific research purposes.

Feature extraction

Considering our aim to employ only predictors that are non-invasive and that are either already routinely assessed or cost-effectively introducible in clinical practice, we decided to focus on information available in our dataset that regard diagnostic subtypes, sociodemographic characteristics, clinical and neuropsychological test scores, cardiovascular risk indexes, and levels of medial temporal lobe brain atrophy in the Hippocampus (HPC), Entorhinal Cortex (ERC), and

Perirhinal Cortex (PRC) as assessed by a clinician-rated Visual Rating Scale (VRS) (Duara, Loewenstein et al. 2008).
Among all the variables related to these domains, some of them were not assessed in all recruited subjects. Variables that had more than 20% of missing values in either the cAD or NC groups were discarded. The following pieces of information were finally used:

- *Sociodemographic characteristics:* gender, age (in years) and years of education calculated by years of schooling and highest degree obtained.
- *MCI subgroups:* Subjects were classified as MCI if they presented subjective memory complaints by the participant and/or or collateral informant, evidence of decline from clinical history and evaluation. All of the MCI patients had a global Clinical Dementia Rating (CDR) score (Morris 1993) of 0.5, and one or more memory measures (including the Hopkins Verbal Learning Test Revised, the Semantic Interference, Logical Memory Delay and Visual Reproduction of the WMS-IV, Trial Making Test, Category Fluency, Letter Fluency and Block Design of the Wechsler Adult Intelligence Scale - Version 3) 1.5 standard deviation or greater below expected normative values were defined as belonging to the amnestic Mild Cognitive Impairment (aMCI) subgroup. MCI subjects with non-memory impairment only were defined as non-amnestic Mild Cognitive Impairment (non-aMCI).
- *PreMCI subgroups:* As defined by Loewenstein and colleagues (Loewenstein, Greig et al. 2012), those individuals who had a global CDR of 0 but had memory or non-memory neuropsychological deficits as described above were diagnosed as Premild Cognitive Impairment - neuropsychological subtype (PreMCI-np). Participants who obtained a global CDR of .5 and had within normal limits performance on neuropsychological testing were classified as Premild Cognitive Impairment - clinical subtype (PreMCI-cl).
- *Clinical scales:* The CDR (Morris 1993) is a 5-point scale (0 = none; 0.5 = very mild, 1 = mild, 2 = moderate, 3 = severe) used to characterize six domains of cognitive and functional performance in Alzheimer disease and related dementias: Memory, Orientation,

Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. The rating is obtained through a semi-structured interview of the patient together with other informants (e.g., family members). The global score was used in the analyses. The memory sum score of a modified informant-based version of CDR (ModCDR-M) was also available and used (range 0-12) (Duara, Loewenstein et al. 2010). The Geriatric Depression Scale (GDS) is a 30-item yes-no self-report assessment used to identify depression in the elderly (Yesavage 1988) and the total score was included in the current analyses (range 0-30).

- Visual Rating Scale for brain atrophy: HPC, ERC, and PRC atrophy levels were assessed with a 0-4 VRS (Duara, Loewenstein et al. 2008). This is an adaptation from the original Scheltens' VRS for the global assessment of medial temporal atrophy (Scheltens, Leys et al. 1992). VRS ratings for HPC, ERC, and PRC were performed in each hemisphere on a Magnetic Resonance Imaging (MRI) image of a standardized coronal slice, perpendicular to the line joining the anterior and posterior commissures, intersecting the mammillary bodies and on adjacent slices. All these 6 VRS measures were separately included as predictors in this study. Ratings are based on a five-point scale: 0 = no atrophy, 1 = minimal atrophy, 2 = mild atrophy, 3 = moderate atrophy, and 4 = severe atrophy. A computer interface provides a library of reference images defining the anatomical boundaries of each brain structure and depicting different levels of atrophy. The whole rating usually takes 5 to 6 minutes per subject (Urs, Potter et al. 2009) and excellent inter-rater (kappa, 0.75 to 0.94) and intra-rater (kappa, 0.84 to 0.94) agreements have been reported (Duara, Loewenstein et al. 2008, Urs, Potter et al. 2009). VRS measures of HPC and ERC have already proved to be predictive of later conversion to Alzheimer's disease in MCI patients (Varon, Barker et al. 2015)
- *Neuropsychological tests:* The Hopkins Verbal Learning Test Revised - Total Recall (HVLTR-R) and Hopkins Verbal Learning Test Revised - Delayed Recall (HVLTR-D) scores (Benedict and Zgaljardic 1998) measuring the verbal learning and memory, the

Semantic Interference Test - Total Retroactive (SIT-RT) and Semantic Interference Test – Total Recognition (SIT-RC) scores (Loewenstein, Acevedo et al. 2004) measuring memory function and interference, and the Trial Making Test - version A (TMT-A) and Trial Making Test - version A (TMT-B), both errors and time (Reitan 1958), measuring visual-motor coordination and attentive functions were considered. Moreover, the Digit-Symbol-Coding Test (DSC), Block Design (Raw Score) and Similarities tests of the Wechsler Adult Intelligence Scale - Version 3 (Wechsler 1997) investigating respectively associative learning, visuospatial function and verbal comprehension, and the Delayed Visual Reproduction Test (DVR) and Logical Memory Test - Immediate Recall (LM-I) and Logical Memory Test - Delayed Recall (LM-D) scores of the WMS-IV (Wechsler 1997) measuring visual and verbal memory were also included.

- *Cardiovascular risk indexes:* Subjects were assessed by physician regarding heart rate, presence or absence of hypertension, high cholesterol levels, diabetes, history of tobacco use, history of myocardial infarction, history of coronary bypass/angioplasty and history of stroke/tia.

Continuous variables were standardized, and categorical variables were coded in order di optimize the number of classes. Categorical cardiovascular risk indexes were re-coded dichotomously and the diagnostic variable was the only polytomous variable, indicating the four diagnostic subgroups (aMCI, non-aMCI, PreMCI-np, PreMCI-cl). In the end, 26 continuous, 9 dichotomous categorical and one four-class categorical features were used. The full list is available in Table 1.
123 subjects have no missing data for all these variables (cAD = 30, 24.39%; NC = 93, 75.61%) and constitute the final sample used in the current study.

Machine learning techniques

Several machine learning procedures exist to solve classification problems. In the current study, we decided to proceed with the following supervised techniques:

- *Elastic Net*: EN is a regression method that adds two types of penalties during the training process. These penalties are the L1 norm of the regression coefficients, as used in LASSO (least absolute shrinkage and selection operator) regression

$$\lambda_1 \sum_{j=1}^{F} |\beta_j|$$

and the L2 norm, as used in ridge regression

$$\lambda_2 \sum_{j=1}^{F} {\beta_j}^2$$

with j indicates the feature, β_j the regression coefficient of the j_{th} feature, and λ_1 and λ_2 are two parameters that define the amount of penalization provided by each of the two terms (Zou and Hastie 2005). The result of including these two penalization terms is a "shrinkage" (i.e., regularization) of the regression coefficients that limit the risk of overfitting, that is when the created algorithm is too good in correctly predicting the cases included in the training sample while having poor performance when used to make prediction in new ones. Moreover, the use of the L1 penalty during training produces also an implicit feature selection, reducing some coefficient to 0 and thus removing some of them form the algorithm. The final predictive model is a logistic regression equation. Thus, the training procedures cannot automatically model non-linear relationships and interactions among predictors, unless polynomials and interactions are "handcrafted" and *a-priori* inserted as features in the model.
- *Elastic Net with polynomial features*: considering what explained above, also EN models including degree three polynomials of the continuous features were trained.
- *Support Vector Machine*: Intuitively, in this algorithm, each case can be viewed as a point in n-dimensional space, where n is

number of features. During the learning process, the linear hyper-plane that optimize the separation of the two classes in such multi-dimensional space is found. New examples are then "plotted" into that space and predicted to belong to a class based on which side they fall on. However, this would allow only to solve so-called linearly separable problems, likewise to what logistic regression can achieve, but SVMs can also perform non-linear classification transforming the original feature space to a higher dimensional space (i.e., creating several new features from the original ones) where the classification problem may better result linearly separable. To perform this transformation in a computationally efficiently manner, the so-called "kernel trick" can be applied, which avoids the explicit transformation that is needed to get linear learning algorithm to learn to perform nonlinear classification. Instead, it enables to operate in an "implicit" feature space without ever computing the coordinate of each case in the new higher dimensional space, but by simply computing the distance of all pairs of cases only considering the original features. In this study, we used the radial basis function (Gaussian) kernel, that is

$$K(\mathrm{x}, \mathrm{x}') = e^{\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)}$$

where x and x' the two feature vectors of two distinct cases and $\|x - x'\|$ is the Euclidean distance between the two (see below for the formula). The kernel parameter σ must be set and requires optimization during the training of the algorithm. Furthermore, also a further C parameter requires optimization. Intuitively, the latter is a regularization parameter that similarly to the λ_2 in EN is useful to improve the generalized performance of the model allowing a trade-off between error in the training sample and model complexity. SVM originally provides only class prediction, with no associated probability. The widely-applied Platt scaling was used to make the model provide probability predictions (Platt 1999). A detailed explanation of SVM and the kernel trick can be found in (Schölkopf, Smola et al. 2002).

- *Gaussian Processes*: GP is a method based on Bayesian theory that can be applied in solving both regression and classification problems, modelling the relationship between the inputs and the outputs following a Bayesian probabilistic approach. A Gaussian process can be viewed as a distribution over functions, and inference consists of applying Bayes' rule to find the posterior function distribution that best approximates the training data. GP used for classification naturally produce probabilistic class predictions. The covariance function matrix of the model can be substituted with a kernel matrix, which represents the counterpart of the "kernel trick" seen before. The radial basis kernel was used also for GP and again this kernel has σ as parameter that requires optimization. A detailed explanation of GP can be found in (Rasmussen 2006).
- *K-Nearest Neighbors*: in the kNN, at first the distances (i.e., the dissimilarity) between a new case and all known examples (i.e., those included in the training set whose output is already known) is calculated. In this analysis, the Euclidean distance was used as distance metric, that is
- $$\sqrt{\sum_{i=1}^{N}(c_i - e_i)^2}$$

  were *c* is the new case, *e* is a known example and *i* is each of the *N* features. To make the prediction, the k less distant examples, also called its nearest neighbors, are taken into account and probabilistic class prediction is performed considering the number of nearest neighbors belonging to each class. K is a hyper-parameter that may take integer values varying from 1 up to the size of the training sample and requires optimization during the training phase.

All analyses were parallelized on a Microsoft® Windows® server equipped with two 6-cores X5650 Intel® Xeon® 2.66 GHz CPUs and were performed

in R (R Core Team 2017), using the implementation of the machine learning techniques available in the caret package (Kuhn 2008).

Cross-validation procedure

All the machine learning techniques used in this study have different so-called hyper-parameters that allow a different tuning of the algorithm during the training process. These are λ_1 and λ_2 in EN and EN-poly, σ and C in SVM, σ in GP, and k in KNN. We trained each model, when possible, with up to 200 random hyper-parameter configurations. Different configurations of these parameters lead to algorithms with different predictive performances. Specifically, we are interested in achieving the best possible performance when the algorithm is applied to new cases that are not part of the training sample.
Considering the small sample size available at this phase, we used cross-validation to provide an estimate of such generalized performance. In cross-validation, the train sample is divided in several folds of cases. Training is iteratively performed with the remaining cases not included in each fold and then the algorithm is tested on the fold cases. Several different cross-validation protocols exist (e.g., n-fold, repeated n-fold, leave-one-out-cross-validation). Recent simulation studies found the rarely applied leave-pair-out cross-validation (LPOCV) protocol to be the best choice when the sample size is limited, being nearly unbiased compared to other commonly applied options such as leave-one-out-cross-validation that instead leads to biased estimate (Parker, Gunter et al. 2007, Airola, Pahikkala et al. 2011). In our study, LPOCV implies to use as folds all possible combinations made of one cAD and one NC. The flaw of LPOCV is its high computational expensiveness. For each attempted hyper-parameter configurations, the training process is performed excluding each defined pair (2790 pairs in the current study) from the training sample and calculating the performance of the algorithm in this left-out pair. Finally, the average performance metric is taken as estimate of the generalized performance of the algorithm created with that particular technique and hyper-parameter configuration.
The performance achieved during the LPOCV procedure will be considered as a first estimate of the performance for the algorithm when

applied to new cases. A test of the model that showed the best LPOCV performance will be performed as a future step using a fully independent dataset. Even if this further investigation is usually lacking for machine learning models developed in the medical field, this will provide a more accurate estimate of the algorithm predictive performance when applied to clinical samples.

Performance metrics

As primary performance metric, the Area Under the Receiving Operating Curve (AUROC) was used. This metric necessitate that the algorithm outputs a single continuous score, in this case in the range 0-1, and then the class prediction is finally made setting a cut-off score (cAD if above or equal to the cut-off score, NC if below). The AUROC value can be interpreted as the probability that a randomly selected cAD subject will receive a higher output score than a randomly selected NC subject, no matter which cut-off is applied to the output score. The AUROC is 0.5 when the algorithm makes predictions at random and 1 in case it is infallible. Considering the LPOCV protocol applied in the analyses, the cross-validated AUROC was calculated with the following formula:

$$\frac{1}{c}\sum_{p=1}^{c}\begin{cases} 1 & if\ \ f_p(x_{cAD,p}) > f_p(x_{NC,p}) \\ 0.5 & if\ f_p(x_{cAD,p}) = f_p(x_{NC,p}) \\ 0 & if\ f_p(x_{cAD,p}) < f_p(x_{NC,p}) \end{cases}$$

were c in the number of LPOCV pairs, f is the output function of the algorithm, $x_{cAD,p}$ is the converter and $x_{NC,p}$ is the non-converter of the each pair. The hyper-parameter configuration for each machine learning technique that produced the best cross-validated AUROC was finally retained. As we could not find in literature any proposed asymptotic procedure to calculate the cross-validated AUROC confidence interval (CI) with the LPOCV protocol, we calculated them with a stratified bootstrap procedure, generating 10000 new subsamples randomly sampling with replacement the original samples and keeping the same frequency of cAD e NC subjects. The distribution of the new 10000 AUROC calculated was used to calculate 95% CI with the bias-corrected and accelerated (BCa) approach (Efron 1987).

The algorithm with the highest performance will be compared to all other algorithms with a paired-sample t-test calculating the standard deviation of the AUROCs difference with 10000 stratified bootstrap resampling, based on what proposed in (Carpenter and Bithell 2000).
Moreover, the cross-validated levels of specificities and balanced accuracy values when sensitivity approached to 0.95, 0.9, 0.85, 0.8, 0.75 were calculated. The cut-off applied to the algorithm output scores was progressively increased starting from 0 and the thresholds providing the closest sensitivity to the aforementioned ones was used to calculate the two other values. The sensitivity and specificity at the best achieved accuracy were also calculated.
To provide distinct predictive performances in the two subpopulations and ease the comparison with previously published models that usually addressed only MCI patients, all performance metrics were also separately calculated in the MCI and PreMCI subsamples. Only the cross-validation pairs containing two MCI and two PreMCI subjects (one cAD and the other NC) were used. Considering that only three converting PreMCI subjects were available, results in the PreMCI subsample should be taken just as preliminary evidence.
The advantage of using AUROC, sensitivity, specificity, and balanced accuracy over other performance metrics (e.g., accuracy, positive predictive value, negative predictive value) is that they are independent from the prevalence of the two outcome classes. Given that the observed rate of conversion to Alzheimer's disease may not be the same in different independent samples, these metrics provide more stable performance estimates and ease the comparison with the performance achieved in other studies.

Feature selection

Training was initially performed including all the 36 features. Only EN and EN-poly automatically operate a selection of features that are finally included in the algorithm. Excluding non-relevant and redundant features and reducing the dimensionality of the algorithm feature-space usually brings to better generalized predictive performance. SVM, GP, KNN, and LR do not automatically operate any feature selection during the training

and so, for these techniques, we re-performed the training and hyper-parameter optimization process with two reduced set of features.
At first, we included only those features selected by the final EN model. Then, we applied a recursive feature elimination (RFE) method with Random Forest as implemented in the rfe function of the caret R package (Kuhn 2008). Detailed description of the algorithm can be found at the following webpage: http://topepo.github.io/caret/recursive-feature-elimination.html. In brief, a Random Forest model is initially trained with all features in each cross-validation fold. Features are ranked according to their importance through a permutation procedure and then the training is re-performed iteratively removing the least ranked feature until when all features have been removed. The optimal number of feature is selected according to the average performance of all cross-validated folds. At the end, the model is trained with the whole sample, features are ranked and those falling in the previously identified optima number of features are retained. As different initial conditions may lead to different final feature subsets, we performed the RFE procedure 100 times with random initialization. We finally included only those features that were selected in more than 50 of the 100 repetitions and we used these to train the SVM, GP, KNN and LR models.
The same paired-sample t-test with bootstrap resampling was also used to test the significance of the change in the LPOCV AUROC achieved applying the two aforementioned feature selection procedures compared to including all the features.

Feature importance

While ranking the importance of features in linear models is straightforward (e.g., in GLM and EN), this is a particularly uneasy task in more complex models (e.g., non-linear kernel SVM and GP). The latter are sometimes referred as black-box models, making hard-to-"impossible" to extract the rules that relate each feature to the outcome. Moreover, different strategy exists for different techniques and a gold-standard procedure has not been defined yet.
To anyhow provide a general ranking of the importance of the predictors, the LPOCV AUROC of each of the 36 features when taken individually

was calculated. This gives a metric of importance for each predictor that is independent from both the applied technique and all other predictors. The 95% CI with the abovementioned stratified bootstrap procedure were also calculated. Feature importance indicated by the LPOCV AUROC were confronted with the selection of features operated by the two feature selection procedures applied in our analyses.

Results

Final analyses required approximately 23 hours of non-stop computations (excluding exploratory and preliminary analyses, and debugging). Descriptive statistics of each feature variables in the cAD and NC groups are reported in Table 1. Statistics of continuous features are reported before the standardization was applied.

Cross-validated predictive performance of algorithms

The cross-validated AUROC for each of the final models is reported in the Table 2 and Fig. 1. SVM, GP and kNN globally achieved better performances then the techniques that cannot model the interaction between the features, i.e., LR and EN. The latter performed generally poorly, even when feature selection strategies were applied to LR and polynomial features were inserted in the EN. LR without feature selection, which was used as reference technique, resulted very poorly performing, being the worst performing model and the sole one showing an AUROC below 0.8 (AUROC = 0.692; C.I. 95% bootstrap=0.598, 0.788).
SVM with the features selected by the RFE procedure is the technique that achieved the highest cross-validated AUROC (AUROC = 0.962; C.I. 95% bootstrap=0.923, 0.987). The results of the paired-sample t-test with stratified bootstrap resampling evidenced that the AUROC of this model was statistically significantly higher ($p<.05$) of all other algorithms, except for the algorithm ranked second (SVM RFE vs GP RFE: $p=.074$). The models achieved high predictive performances also when the two subsample were considered separately, although lower in the MCI subsample (AUROC=0.914; C.I. 95% bootstrap=0.822, 0.975) and very

high in the PreMCI subsample (AUROC=0.994; C.I. 95% bootstrap=0.932, 1).
The cross-validated levels of specificities and balanced accuracy values when sensitivity approached 0.95, 0.9, 0.85, 0.8, 0.75, as much as the sensitivity and specificity at the best achieved accuracy are reported in Table 3. Considering the whole sample of both MCI and PreMCI subjects, the best achieved cross-validated balanced accuracy is 0.913 (sensitivity = 0.956, specificity = 0.871). Again, performances were still high but lower in magnitude in the MCI subsample, with a best balanced accuracy of 0.874 (sensitivity = 0.880, specificity = 0.867). Instead, preliminary results in the PreMCI subsample presented very high performances, with a best balanced accuracy of 0.980 (sensitivity = 1, specificity = 0.960).

Efficacy of feature selection procedures

The features selected by the EN model with the best hyper-parameter configuration and the RFE procedure are also specified in Table 1. The RFE procedure used in this study resulted effective in identifying a relevant subset of the initial features, leading for all the techniques to a significant improvement of the cross-validated performances compared to the use of all features (SVM vs SVM RFE: $p = .015$; GP vs GP RFE: $p = .023$; kNN vs kNN RFE: $p=.048$; LR vs LR RFE: $p<.001$). Moreover, also the models ranked second and third were GP and kNN with the features selected by the RFE procedure and they both achieved an AUROC higher than 0.9.
Instead, the approach of using the features selected by the EN model was not particularly efficacious, leading to not statistically significant improvements in GP, kNN, and LR and even leading to a reduced performance in SVM.

Feature importance

The LPOCV AUROC of each of the 36 features is reported in Table 4, ranked from the highest to the lowest AUROC, and in Fig. 2, subdivided based on their type (i.e., sociodemographic, diagnosis, clinical, VRS, neuropsychological tests and of cardiovascular risk indexes).

The sociodemographic features had poor predictive capability. All their AUROC resulted below 0.65 and only age achieved statistical significance (lower bound of the 95% C.I. higher than 0.5). As a matter of facts, neither the EN model nor the RFE procedures selected any of the sociodemographic features to be included in the models.
The baseline diagnosis (i.e., aMCI, non-aMCI, PreMCI-np, and PreMCI-cl) resulted instead quite predictive, with an AUROC of 0.759. This is again in accordance with both the feature selection procedures that identified these features as to be retained.
Among the clinical scales, only the ModCDR-M score resulted with both a significant and relevant cross-validated AUROC (AUROC = 0.730), being the sole selected by both the feature selection procedures. The global CDR score, although resulting with a statistically significant AUROC, had an AUROC very small in magnitude (AUROC = 0.559).
The AUROC of the six VRS scores ranged from 0.761 (right ERC atrophy) to 0.647 (the left PRC atrophy). The left PRC atrophy score was the sole not selected by the RFE procedure while all VRS scores were included in the final EN model.
Among the fourteen neuropsychological test scores, the HVLTR-R and HVLTR-D scores, the SIT-RT and SIT-RC scores, LM-I and LM-D scores of the Weschler Memory Scale – Fourth Edition (WMS-IV) resulted the tests with the highest predictive performances (all AUROC above 0.750) and these were all selected by both the feature selection procedures. The DVR score of the WMS-IV also resulted able to provide statistically significant although less precise prediction of conversion (AUROC = 0.718), as much as TMT-A and TMT-B errors (AUROC ranging between 0.6 and 0.7). Of these, both time and errors of the TMT-B resulted included also in the final EN model, while the RFE procedure selected only TMT-B errors.
Finally, among the cardiovascular risk features, only history of stroke/tia and history of coronary bypass/angioplasty were found to have an AUROC statistically significant and higher than 0.6. Interestingly, the selection of these features by the two feature selection procedures resulted quite different from this evidence. The final EN model didn't include any of the cardiovascular risk features, while the RFE selected history of myocardial infarction and heart rate, which had a non-

significant LPOCV AUROC, and not history of myocardial infarction and history of coronary bypass/angioplasty.

Discussion

The current study represents the first step in the development of a novel machine-learning algorithm for the identification of three-year conversion to Alzheimer's disease in subjects with either MCI or PreMCI. Such an algorithm finally aims to be efficiently applicable in clinical practice, thus achieving high accuracy and to be based on predictors that can be easily and effectively assessed in clinical settings.
The algorithms developed in this study promise to fulfill both these requirements. We employed only predictors based on sociodemographic characteristics, clinical and neuropsychological tests, cardiovascular risk indexes, and level of brain atrophy as assessed by clinicians through the VRS from structural MRI images. With these pieces of information, our best algorithm achieved a global cross-validated AUROC higher than 0.96, with a AUROC higher than 0.91 also in the MCI subsample. This indicates that our best algorithm already outperforms the clear majority of the several previously proposed algorithms. Furthermore, to the best of our knowledge, this is the only available predictive model that was developed for subjects at a PreMCI stage, showing very high preliminary performance (AUROC > 0.99) also in the PreMCI subgroup.

Translation to clinical practice

Among all the algorithms we developed, the one which showed the best performance was the SVM with radial-basis function kernel that included only the features selected via the RFE procedure. Regarding the MCI subsample, roughly 88% of specificity and 87% sensitively are the levels that resulted maximizing the overall cross-validated balanced accuracy (87%). Also, we found results of a nearly perfect identification of cAD in the PreMCI subsample (cross-validated accuracy = 98%), although these should be considered preliminary as we only had three cAD

PreMCI subjects in our sample. Further testing in independent clinical samples would finally confirm these results.
The predictive capabilities achieved by this model would make its application useful in clinical practice as much as in clinical trials, representing a relevant improvement in the current possibility to identify only those subjects truly at risk of converting to Alzheimer's disease. Moreover, it would be possible to further optimize the desired levels of specificity and sensitivity according to the cost associated in predicting false positives and negatives.
We achieved the obtained results employing only information that can be collected in routine clinical practice. All the measures we used as predictors are non-invasive and can be easily introduced in any clinical center without requiring any particular difficulty or the purchase of non-standardly available equipment. All the neuropsychological tests do not necessitate any intensive training and can be administered by a technician under the supervision of a neuropsychologist. Moreover, the availability of machines for structural MRI and the ease of using VRS is fast and easily adoptable thanks to the availability of a software with reference images that guide the clinician during the rating, providing training for the relatively uninitiated radiologist, neurologist, or any other interested rater (Urs, Potter et al. 2009). The VRS overcomes the issue of MRI data obtained from different machines, which are usually non-automatically comparable. All the remaining information we considered, such as socio-demographic, clinical, and cardiovascular risk, can be readily collected during neurological interviews.

Comparisons with other available machine learning algorithms

Several machine learning algorithms have been previously proposed to predict the MCI to Alzheimer's disease conversion. Among those that used only baseline information and make a prediction of conversion in about three years, we could identify only a few achieving performances similar or superior to ours, and they are reported in Table 5.
Specifically, five studies evidenced superior performances. The algorithm proposed by Argwal and colleagues (Agarwal, Ghanty et al. 2015) uses a selection of blood plasma proteins as sole predictors. This is a very interesting result as their model uses information from a different

domain and it may be partially complementary to the features we used. Also, the prediction is entirely based on the analysis of a single blood sample and even if the assessment of such protein blood levels is not currently clinical routine, it requires a very little invasive procedure and may be develop so to result cost-effectively adoptable in clinical practice. However, these results come from a small training sample and further investigation is necessary to evidence the soundness of such promising results.

Three further algorithms have been developed based on structural MRI data: those proposed by Minhas and colleagues (Minhas, Khanum et al. 2017), and Plant and colleagues (Plant, Teipel et al. 2010) were trained and cross-validated in very small samples, respectively of 13 and 24 MCI subjects, while Long and colleagues (Long, Chen et al. 2017) used a larger sample (n=227). All these algorithms showed very high cross-validated performance. However, they directly use structural MRI data and considering the difficulties of employing together data coming from different scanners (Teipel, Reuter et al. 2011), this may place a barrier to an efficient dissemination of such algorithms into clinical practice.

Finally, also Hojjati and collegues proposed an algorithm (Hojjati, Ebrahimzadeh et al. 2017) with high predictive accuracy based on resting state functional MRI data. If the availability of MRI machines in clinical setting is quite common nowadays, functional MRI is still mainly used in research settings. Thus, such algorithm may currently result difficulty applicable in clinical practice.

Four additional studies proposed algorithm with performances similar to ours. Three studies employed predictors that may not allow an easy translation to clinical practice: Dukart and colleagues (Moradi, Pepe et al. 2015) used structural MRI data, Dukart and colleagues (Dukart, Sambataro et al. 2016) both structural MRI and fludeoxyglucose positron emission tomography data, and Apostolova and colleagues (Apostolova, Hwang et al. 2014) cerebrospinal fluid p-tau protein levels.

Instead, Clark and colleagues (Clark, Kapur et al. 2014) used only sociodemographic, clinical, and neuropsychological test scores, achieving high cross-validated performances although inferior to those achieved by our best model.

Considering this evidence, our and Clark's algorithms are the only two currently available that achieved a relevant predictive performance using

only predictors that may be easily assessed in nowadays clinical practice, with our algorithm that seems to outperform the Clark's ones. As we used different predictors than those employed in Clark's algorithms (i.e., they didn't use brain atrophy levels assessed via the VRS but included the scores of a novel semantic fluency word lists test), it would be of great interest to investigate in the next steps if adding such predictors to our features would bring to a further increase in the predictive performance of our algorithms.

Importance of predictors

While sociodemographic and cardiovascular risk were not particularly predictive, memory and brain atrophy seem to be the most relevant for the prediction of Alzheimer's disease conversion. The HVLTR, SIT, and LM tests were identified as the most relevant cognitive measures by all feature selection and importance procedures, and they all assess different aspects of memory. The ModCDR-M score was also suggested as a particularly relevant feature. The important role of memory functioning as predictor was somehow expected considering previous findings (Loewenstein, Acevedo et al. 2007) and that memory deficits are the core clinical characteristics that defines Alzheimer's disease. Also, the evidence of an important role of brain atrophy is in line with previous evidence (Li, Tan et al. 2016) as well as several other studies which developed highly performing machine learning algorithms starting from structural MRI data, alone (e.g., Plant, Teipel et al. 2010) or in combination with neuropsychological test scores (e.g., Moradi, Pepe et al. 2015, Minhas, Khanum et al. 2017). Memory deterioration and brain atrophy may begin years before a full-blown Alzheimer's disease diagnosis can be made and a proper set of sensible measures can allow to promptly identify them. Our study further suggests that machine learning techniques have the potential to exploit such information to early identify those subjects in which the onset of the pathophysiological processes leading to Alzheimer's disease has been occurring.

Limitations

Our study has some potential limitations that should be taken into account. We used cross-validation as validation procedure but further

testing in an independent sample of new cases has not been performed yet. However, nearly all the algorithms proposed to make a MCI-to-Alzheimer's disease prediction currently lack such further testing. Furthermore, the sample we used to train the algorithm was limited in size and included only three cAD PreMCI. Thus, the performance estimate obtained for the PreMCI should be considered as very preliminary and requires further investigation.

Also, we applied only some of the many machine learning as well as feature selection procedures available. Although we have already reached good results, there is no guarantee that other machine learning procedures and other subsets of features would allow to achieve even better predictive accuracy.

Moreover, all subjects of our sample were recruited in the same abovementioned clinical centers. The population referring to these might have peculiar characteristics and algorithms might perform less well in different MCI and PreMCI populations. Also, both the features and subjects we finally included were selected from a larger set of available variables and subjects according to the lack of missing values. Their occurrence in such excluded variables and subjects may be due to reasons that are beyond mere randomness, potentially limiting the representativeness of our feature set and train sample and thus leading to biases in our algorithms.

Given these current issues, we plan to test the performance in a new sample of MCI and PreMCI subjects currently in a longitudinal study in Miami, currently in its third year, as well as to try new procedures for further optimization.

Another potential shortcoming is the complexity of providing a clear explanation of the role that each feature plays in the prediction. While a first basic approach has been attempted in this study, more strategies will be applied while proceeding in the next phases with larger samples. A better interpretability of the model will help both in gaining further understandings of how these variables are related to the development of Alzheimer's disease and in generating more trust towards the application of model by clinicians as much as patients.

Conclusion

In conclusion, we used supervised machine learning techniques to develop algorithms able to identify which subjects with PreMCI and MCI will convert to Alzheimer's disease in the following three years. As the opportunity of an efficient clinical translation was one of the main goals motivating our study, we used predictors based only on sociodemographic characteristics, clinical tests, cognitive measures, cardiovascular risk indexes, and level of brain atrophy as assessed by clinicians through the VRS from structural MRI images. We promisingly achieved high predictive performance, among the very best of the many algorithms available in literature and the best achieved so far using only information easily assessable in clinical practice. Considering these results, we plan to proceed in further testing and optimization in other independent and larger samples as to reach the level of reliability necessary for an actual applicability.

Acknowledgments: this study was supported by the National Institute on Aging, United States Department of Health and Human Services (grant: 1P50AG025711-05, R01AG047649-01A1, P50 AG047726602).

References

Agarwal, S., P. Ghanty and N. R. Pal (2015). "Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease." Bioinformatics **31**(15): 2505-2513.

Airola, A., T. Pahikkala, W. Waegeman, B. De Baets and T. Salakoski (2011). "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve." Computational Statistics & Data Analysis **55**(4): 1828-1844.

American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision). 4 ed. Washington, DC, Amer Psychiatric Pub.

Apostolova, L. G., K. S. Hwang, O. Kohannim, D. Avila, D. Elashoff, C. R. Jack, Jr., L. Shaw, J. Q. Trojanowski, M. W. Weiner, P. M.

Thompson and I. Alzheimer's Disease Neuroimaging (2014). "ApoE4 effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer's disease." Neuroimage Clin **4**: 461-472.

Benedict, R. H. and D. J. Zgaljardic (1998). "Practice effects during repeated administrations of memory tests with and without alternate forms." J Clin Exp Neuropsychol **20**(3): 339-352.

Breitner, J. C. (2014). "Mild cognitive impairment and progression to dementia: new findings." Neurology **82**(4): e34-35.

Brooks, L. G. and D. A. Loewenstein (2010). "Assessing the progression of mild cognitive impairment to Alzheimer's disease: current trends and future directions." Alzheimers Res Ther **2**(5): 28.

Carpenter, J. and J. Bithell (2000). "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians." Stat Med **19**(9): 1141-1164.

Chao, L. L., S. G. Mueller, S. T. Buckley, K. Peek, S. Raptentsetseng, J. Elman, K. Yaffe, B. L. Miller, J. H. Kramer, C. Madison, D. Mungas, N. Schuff and M. W. Weiner (2010). "Evidence of neurodegeneration in brains of older adults who do not yet fulfill MCI criteria." Neurobiol Aging **31**(3): 368-377.

Clark, D. G., P. Kapur, D. S. Geldmacher, J. C. Brockington, L. Harrell, T. P. DeRamus, P. D. Blanton, K. Lokken, A. P. Nicholas and D. C. Marson (2014). "Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease." Cortex **55**: 202-218.

Cooper, C., A. Sommerlad, C. G. Lyketsos and G. Livingston (2015). "Modifiable predictors of dementia in mild cognitive impairment: a systematic review and meta-analysis." Am J Psychiatry **172**(4): 323-334.

Duara, R., D. A. Loewenstein, M. T. Greig-Custo, A. Raj, W. Barker, E. Potter, E. Schofield, B. Small, J. Schinka, Y. Wu and H. Potter (2010). "Diagnosis and staging of mild cognitive impairment, using a modification of the clinical dementia rating scale: the mCDR." Int J Geriatr Psychiatry **25**(3): 282-289.

Duara, R., D. A. Loewenstein, E. Potter, J. Appel, M. T. Greig, R. Urs, Q. Shen, A. Raj, B. Small, W. Barker, E. Schofield, Y. Wu and H. Potter

(2008). "Medial temporal lobe atrophy on MRI scans and the diagnosis of Alzheimer disease." Neurology **71**(24): 1986-1992.

Dukart, J., F. Sambataro and A. Bertolino (2016). "Accurate Prediction of Conversion to Alzheimer's Disease using Imaging, Genetic, and Neuropsychological Biomarkers." J. Alzheimers. Dis. **49**(4): 1143-1159.

Efron, B. (1987). "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association **82**(397): 171-185.

Forlenza, O. V., B. S. Diniz, A. L. Teixeira, F. Stella and W. Gattaz (2013). "Mild cognitive impairment. Part 2: Biological markers for diagnosis and prediction of dementia in Alzheimer's disease." Rev Bras Psiquiatr **35**(3): 284-294.

Hojjati, S. H., A. Ebrahimzadeh, A. Khazaee, A. Babajani-Feremi and I. Alzheimer's Disease Neuroimaging (2017). "Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM." J. Neurosci. Methods **282**: 69-80.

International, A. s. D. (2016). World Alzheimer Report 2016. Improving healthcare for people

living with dementia.

Kang, J. H., M. Korecka, J. B. Toledo, J. Q. Trojanowski and L. M. Shaw (2013). "Clinical utility and analytical challenges in measurement of cerebrospinal fluid amyloid-beta(1-42) and tau proteins as Alzheimer disease biomarkers." Clin Chem **59**(6): 903-916.

Klunk, W. E. (2011). "Amyloid imaging as a biomarker for cerebral beta-amyloidosis and risk prediction for Alzheimer dementia." Neurobiol Aging **32 Suppl 1**: S20-36.

Kuhn, M. (2008). "Building Predictive Models in R Using the caret Package." Journal of Statistical Software **28**(5).

Li, J. Q., L. Tan, H. F. Wang, M. S. Tan, L. Tan, W. Xu, Q. F. Zhao, J. Wang, T. Jiang and J. T. Yu (2016). "Risk factors for predicting progression from mild cognitive impairment to Alzheimer's disease: a systematic review and meta-analysis of cohort studies." J Neurol Neurosurg Psychiatry **87**(5): 476-484.

Loewenstein, D. A., A. Acevedo, J. Agron and R. Duara (2007). "Vulnerability to proactive semantic interference and progression to

dementia among older adults with mild cognitive impairment." Dement Geriatr Cogn Disord **24**(5): 363-368.

Loewenstein, D. A., A. Acevedo, C. Luis, T. Crum, W. W. Barker and R. Duara (2004). "Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia." J Int Neuropsychol Soc **10**(1): 91-100.

Loewenstein, D. A., R. E. Curiel, R. Duara and H. Buschke (2017). "Novel Cognitive Paradigms for the Detection of Memory Impairment in Preclinical Alzheimer's Disease." Assessment: 1073191117691608.

Loewenstein, D. A., M. T. Greig, J. A. Schinka, W. Barker, Q. Shen, E. Potter, A. Raj, L. Brooks, D. Varon, M. Schoenberg, J. Banko, H. Potter and R. Duara (2012). "An investigation of PreMCI: subtypes and longitudinal outcomes." Alzheimers Dement **8**(3): 172-179.

Long, X., L. Chen, C. Jiang, L. Zhang and I. Alzheimer's Disease Neuroimaging (2017). "Prediction and classification of Alzheimer disease based on quantification of MRI deformation." PLoS One **12**(3): e0173372.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price and E. M. Stadlan (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology **34**(7): 939-944.

Minhas, S., A. Khanum, F. Riaz, A. Alvi and S. A. Khan (2017). "A Nonparametric Approach for Mild Cognitive Impairment to AD Conversion Prediction: Results on Longitudinal Data." IEEE J Biomed Health Inform **21**(5): 1403-1410.

Moradi, E., A. Pepe, C. Gaser, H. Huttunen, J. Tohka and I. Alzheimer's Disease Neuroimaging (2015). "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects." Neuroimage **104**: 398-412.

Morris, J. C. (1993). "The Clinical Dementia Rating (CDR): current version and scoring rules." Neurology **43**(11): 2412-2414.

Parker, B. J., S. Gunter and J. Bedo (2007). "Stratification bias in low signal microarray studies." BMC Bioinformatics **8**: 326.

Petersen, R. C., J. E. Parisi, D. W. Dickson, K. A. Johnson, D. S. Knopman, B. F. Boeve, G. A. Jicha, R. J. Ivnik, G. E. Smith, E. G. Tangalos, H. Braak and E. Kokmen (2006). "Neuropathologic features of amnestic mild cognitive impairment." Arch Neurol **63**(5): 665-672.

Plant, C., S. J. Teipel, A. Oswald, C. Bohm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel and M. Ewers (2010). "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease." Neuroimage **50**(1): 162-174.

Platt, J. C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." ADVANCES IN LARGE MARGIN CLASSIFIERS: 61-74.

R Core Team (2017). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.

Rasmussen, C. E. (2006). "Gaussian processes for machine learning."

Reitan, R. M. (1958). "Validity of the Trail Making Test as an Indicator of Organic Brain Damage." Percept. Mot. Skills **8**(3): 271-276.

Roberts, R. O., D. S. Knopman, M. M. Mielke, R. H. Cha, V. S. Pankratz, T. J. H. Christianson, Y. E. Geda, B. F. Boeve, R. J. Ivnik, E. G. Tangalos, W. A. Rocca and R. C. Petersen (2014). "Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal." Neurology **82**(4): 317-325.

Samuel, A. L. (1959). "Some studies in machine learning using the game of checkers." IBM Journal of research and development **3**(3): 210-229.

Scheltens, P., D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters and J. Valk (1992). "Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates." J Neurol Neurosurg Psychiatry **55**(10): 967-972.

Schölkopf, B., A. J. Smola and F. Bach (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press.

Sperling, R. and K. Johnson (2013). "Biomarkers of Alzheimer disease: current and future applications to diagnostic criteria." Continuum (Minneap Minn) **19**(2 Dementia): 325-338.

Szeto, J. Y. and S. J. Lewis (2016). "Current Treatment Options for Alzheimer's Disease and Parkinson's Disease Dementia." Curr Neuropharmacol **14**(4): 326-338.

Teipel, S. J., S. Reuter, B. Stieltjes, J. Acosta-Cabronero, U. Ernemann, A. Fellgiebel, M. Filippi, G. Frisoni, F. Hentschel, F. Jessen, S. Kloppel, T. Meindl, P. J. Pouwels, K. H. Hauenstein and H. Hampel (2011). "Multicenter stability of diffusion tensor imaging measures: a European clinical and physical phantom study." Psychiatry Res **194**(3): 363-371.

Urs, R., E. Potter, W. Barker, J. Appel, D. A. Loewenstein, W. Zhao and R. Duara (2009). "Visual rating system for assessing magnetic resonance images: a tool in the diagnosis of mild cognitive impairment and Alzheimer disease." J Comput Assist Tomogr **33**(1): 73-78.

Van Cauwenberghe, C., C. Van Broeckhoven and K. Sleegers (2016). "The genetic landscape of Alzheimer disease: clinical implications and perspectives." Genet Med **18**(5): 421-430.

van Rossum, I. A., S. Vos, R. Handels and P. J. Visser (2010). "Biomarkers as predictors for conversion from mild cognitive impairment to Alzheimer-type dementia: implications for trial design." J Alzheimers Dis **20**(3): 881-891.

Varon, D., W. Barker, D. Loewenstein, M. Greig, A. Bohorquez, I. Santos, Q. Shen, M. Harper, T. Vallejo-Luces, R. Duara and I. Alzheimer's Disease Neuroimaging (2015). "Visual rating and volumetric measurement of medial temporal atrophy in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort: baseline diagnosis and the prediction of MCI outcome." Int J Geriatr Psychiatry **30**(2): 192-200.

Wechsler, D. (1997). WAIS-III: Administration and scoring manual: Wechsler adult intelligence scale, Psychological Corporation.

Wechsler, D. (1997). WMS-III: Wechsler memory scale administration and scoring manual, Psychological Corporation.

Yesavage, J. A. (1988). "Geriatric Depression Scale." Psychopharmacol Bull **24**(4): 709-711.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) **67**(2): 301-320.

Table 1. Descriptive statistics.

Continuous Predictors	Non-converters		Converters		Selected by		
	Mean	S.D.	Mean	S.D.	EN	RFE %	RFE (>50%)
Age	73.57	6.19	76.53	6.27		22%	
Years of education	13.23	3.15	14.70	4.59		2%	
Body-Mass Index	69.27	8.35	71.47	8.17		2%	
Modified Clinical Dementia Rating Scale, Memory Sum Score	1.86	1.32	2.92	1.27	x	99%	x
Geriatric Depression Scale	2.28	2.27	2.17	2.82		1%	
Right hippocampus atrophy (VRS)	0.90	0.87	1.77	1.17	x	100%	x
Left hippocampus atrophy (VRS)	0.87	0.82	1.70	1.06	x	100%	x
Right entorhinal cortex atrophy (VRS)	0.54	0.76	1.53	1.17	x	100%	x
Left entorhinal cortex atrophy (VRS)	0.65	0.82	1.50	1.23	x	100%	x
Right perirhinal cortex atrophy (VRS)	0.53	0.70	1.27	1.20	x	100%	x
Left perirhinal cortex atrophy (VRS)	0.51	0.72	1.10	1.19	x	34%	
Hopkins Verbal Learning Test Revised, Total Recall	21.48	5.31	16.20	4.71	x	96%	x

Hopkins Verbal Learning Test Revised, Delayed Recall	6.99	3.09	3.10	2.94	x	100%	x
Digit-Symbol-Coding Score (WAIS-III)	11.17	2.64	10.47	2.43		17%	
Block Design Test, Raw Score (WAIS-III)	26.77	8.54	23.67	7.33		4%	
Semantic Interference Test, Retroactive Total Score	22.89	3.80	16.87	5.88	x	100%	x
Semantic Interference Test, Recognition Total Score	26.43	3.94	21.03	5.75	x	100%	x
Delayed Visual Reproduction (WMS-IV)	15.74	9.60	8.83	9.46	x	12%	
Logical Memory, Immediate Recall Score (WMS-IV)	10.95	3.95	6.73	3.27	x	100%	x
Logical Memory, Delayed Recall Score (WMS-IV)	8.62	4.49	3.97	3.16	x	100%	x
Trial Making A, Time (in seconds)	39.89	13.99	49.63	18.67		46%	
Trial Making A, Errors	0.17	0.43	0.10	0.31		7%	
Trial Making B, Time (in seconds)	132.18	70.32	167.90	73.45	x	31%	
Trial Making B, Errors	1.68	3.65	1.90	1.81	x	63%	x
Heart Rate (in beats-per-minute)	27.91	5.39	25.97	3.73		92%	x

Categorical Predictors		Non-converters		Converters		Selected by		
		N	%	N	%	EN	RFE %	RFE (>50%)
Gender	Male	41	44.09%	12	40.00%		12%	
	Female	52	55.91%	18	60.00%			
Clinical Dementia Rating Scale. Global Score	0	12	12.90%	0	0.00%		38%	
	0.5	81	87.10%	30	100%			
Hypercholesterolemia	No	34	36.56%	7	23.33%		1%	
	Yes	59	63.44%	23	76.67%			
Hypertension	No	42	45.16%	16	53.33%		2%	
	Yes	51	54.84%	14	46.67%			
Diabetes	No	79	84.95%	26	86.67%		2%	
	Yes	14	15.05%	4	13.33%			
History of tobacco use	No	87	93.55%	29	96.67%		1%	
	Yes	6	6.45%	1	3.33%			
History of myocardial infarction	No	88	94.62%	5	16.67%		100%	x
	Yes	26	27.96%	4	13.33%			
History of coronary bypass/angioplasty	No	80	86.02%	26	86.67%		41%	
	Yes	13	13.98%	4	13.33%			
History of stroke/tia	No	80	86.02%	25	83.33%		6%	
	Yes	13	13.98%	5	16.67%			
Diagnosis	aMCI	27	29.03%	25	83.33%	x	100%	x
	non-aMCI	7	7.53%	2	6.67%			
	PreMCI-cl	31	33.33%	1	3.33%			
	PreMCI-np	28	30.11%	2	6.67%			

S.D. = Standard Deviation; EN = Elastic Net; RFE Recursive Feature Elimination; N = number of subjects; aMCI = amnestic Mild Cognitive Impairment; non-aMCI = non-amnestic Mild Cognitive Impairment; PreMCI-cl = Premild Cognitive Impairment - clinical subtype; PreMCI-np = Premild Cognitive Impairment - neuropsychological subtype; WAIS-III = Wechsler Adult Intelligence Scale - Version 3; WMS-IV = Weschler Memory Scale – Fourth Edition; VRS = Visual Rating Scale.

Table 2. Leave-Pair-Out-Cross-Validation AUROC of the final algorithms.

Method	LPOCV AUROC	CI 95% (Bootstrap)		Comparison with the Best Algorithm	
				t (Bootstrap)	p
SVM (features selected by RFE)	0.962	0.923	0.987	-	-
GP (features selected by RFE)	0.935	0.886	0.970	1.802	0.074
kNN (features selected by RFE)	0.916	0.859	0.958	2.216	0.029
SVM	0.910	0.846	0.956	2.457	0.015
GP (features selected by EN)	0.900	0.832	0.948	2.543	0.012
GP	0.899	0.833	0.944	3.052	0.003
kNN (features selected by EN)	0.894	0.825	0.945	2.775	0.006
SVM (features selected by EN)	0.889	0.822	0.939	3.002	0.003
EN	0.889	0.805	0.933	2.828	0.005
kNN	0.886	0.814	0.941	2.837	0.005
EN-poly	0.878	0.816	0.942	3.159	0.002
LR (features selected by RFE)	0.832	0.733	0.909	3.339	0.001
LR (features selected by EN)	0.827	0.733	0.899	3.280	0.001
LR	0.692	0.598	0.788	5.714	<0.001

AUROC = Area Under the Receiving Operating Curve; LPOCV = leave-pair-out cross-validation; CI = Confidence Interval; EN = Elastic Net; EN-poly = Elastic Net with Polynomial features; GP = Gaussian Processes; kNN = k-Nearest Neighbors; LR = Logistic Regression; RFE = recursive feature elimination; SVM = Support Vector Machine.

Table 3. Performance metrics of the best model. SVM with features selected by RFE.

	Performance level (target)	Sensitivity	Specificity	Balanced Accuracy
Whole Sample (AUROC = 0.962)	Sensitivity of .95	0.950	0.874	0.912
	Sensitivity of .90	0.900	0.906	0.903
	Sensitivity of .85	0.850	0.938	0.894
	Sensitivity of .80	0.800	0.951	0.876
	Sensitivity of .75	0.750	0.957	0.853
	Best Balanced Accuracy	0.956	0.871	0.913
Only MCI (AUROC = 0.914)	Sensitivity of .95	0.951	0.734	0.843
	Sensitivity of .90	0.901	0.822	0.862
	Sensitivity of .85	0.851	0.877	0.864
	Sensitivity of .80	0.801	0.882	0.842
	Sensitivity of .75	0.751	0.883	0.817
	Best Balanced Accuracy	0.880	0.867	0.874
Only PreMCI (AUROC = 0.994)	Sensitivity of .95	0.955	0.960	0.958
	Sensitivity of .90	0.904	0.966	0.935
	Sensitivity of .85	0.853	0.966	0.910
	Sensitivity of .80	0.802	0.966	0.884
	Sensitivity of .75	0.751	0.966	0.859
	Best Balanced Accuracy	1.000	0.960	0.980

AUROC = Area Under the Receiving Operating Curve; MCI = Mild Cognitive Impairment; PreMCI = Premild Cognitive Impairment.

Table 4. Feature importance.

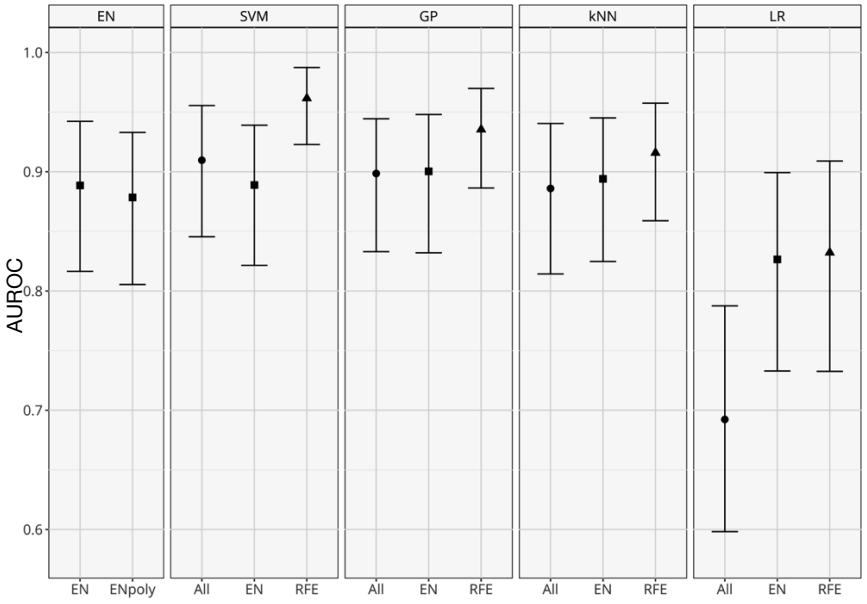
Feature	LPOCV AUROC	CI 95% (Bootstrap)		Statistical significance
Hopkins Verbal Learning Test Revised, Delayed Recall	0.812	0.722	0.887	x
Semantic Interference Test, Retroactive Total Score	0.810	0.715	0.888	x
Logical Memory, Delayed Recall Score (WMS-IV)	0.794	0.703	0.871	x
Logical Memory, Immediate Recall Score (WMS-IV)	0.789	0.688	0.868	x
Hopkins Verbal Learning Test Revised, Total Recall	0.779	0.687	0.858	x
Semantic Interference Test, Recognition Total Score	0.778	0.664	0.868	x
Right entorhinal cortex atrophy (VRS)	0.761	0.654	0.853	x
Left hippocampus atrophy (VRS)	0.732	0.629	0.827	x
Modified Clinical Dementia Rating Scale, Memory Sum Score	0.730	0.627	0.824	x
Right hippocampus atrophy (VRS)	0.721	0.614	0.816	x
Delayed Visual Reproduction (WMS-IV)	0.718	0.599	0.818	x
Left entorhinal cortex atrophy (VRS)	0.708	0.597	0.809	x
Right perirhinal cortex atrophy (VRS)	0.691	0.585	0.794	x
Trial Making A, Time (in seconds)	0.664	0.554	0.765	x
Trial Making B, Time (in seconds)	0.659	0.543	0.766	x
Left perirhinal cortex atrophy (VRS)	0.647	0.538	0.753	x
History of stroke/tia	0.630	0.686	0.580	x
Trial Making B, Errors	0.625	0.515	0.733	x

Table 5. Comparison with previous algorithms for MCI subjects with comparable of superior performance

Reference	Follow-up Period	Predictors	Machine Learning Technique	Validation Protocol	Sample Size	AUROC	Specificity	Sensitivity	Balanced Accuracy
Our best algorithm	3 years	Socio-demographic, clinical, neuropsychological, clincian-rated brain atrophy, cadiovascular risk scores	SVM with radial-basis function kernel	Leave-pair-out (>50%) cross-validation	61	0.914	0.880	0.867	0.874
(Plant, Teipel et al. 2010)	Approximately 2.5 years	structural MRI	Linear SVM	Leave-one-out cross-validation	24	n.a.	0.889	1.000	0.945
			Bayesian Classifier	Leave-one-out cross-validation		n.a.	0.778	1.000	0.889
			Voting Feature Intervals	Leave-one-out cross-validation		n.a.	1.000	0.933	0.967
(Hojjati, Ebrahimzadeh et al. 2017)	3 years	Resting-state functional MRI	Linear SVM	Repeated-9-fold cross-validation	80	0.949	0.901	0.832	n.a.
(Minhas, Khanum et al. 2017)	3 years	Structural MRI, Neuropsychological tests	Novel non-parametric approach	Leave-one-out cross-validation	13	n.a.	0.923	0.875	0.899
(Agarwal, Ghanty et al. 2015)	5/6 years	blood plasma proteins	Radial-basis function network	repeated 5-fold cross-validation (10 repetitions)	47	n.a.	>0.9	>0.95	>0.925
(Long, Chen et al. 2017)	3 years	Structural MRI	Linear SVM	10-fold cross-validation	227	0.932	0.909	0.863	n.a.
(Dukart, Sambataro et al. 2016)	At least 2 years	APOE typization, FDG-PET, and sturctural MRI	Naïve Bayes	Train dataset: AD + controls; Test dataset: MCI	192	0.840	0.861	0.875	0.868
(Moradi, Pepe et al. 2015)	3 years	structural MRI, neuropsychological measures and age	Semi-supervized approach and Random Forest	10-fold cross-validation	264	0.902	0.740	0.870	0.805
(Apostolova, Hwang et al. 2014)(only ApoE4-negative subjects)	3 years	CSF p-tau, Education, Sex	SVM with radial-basis function kernel	Leave-one-out cross-validation	83	0.890	n.a.	n.a.	n.a.
(Clark, Kapur et al. 2014)	At least 1 year	Socio-demographic, clinical, and neuropsychological test scores	Random Forest	10-fold cross-validation	80	0.880	0.900	0.780	0.840

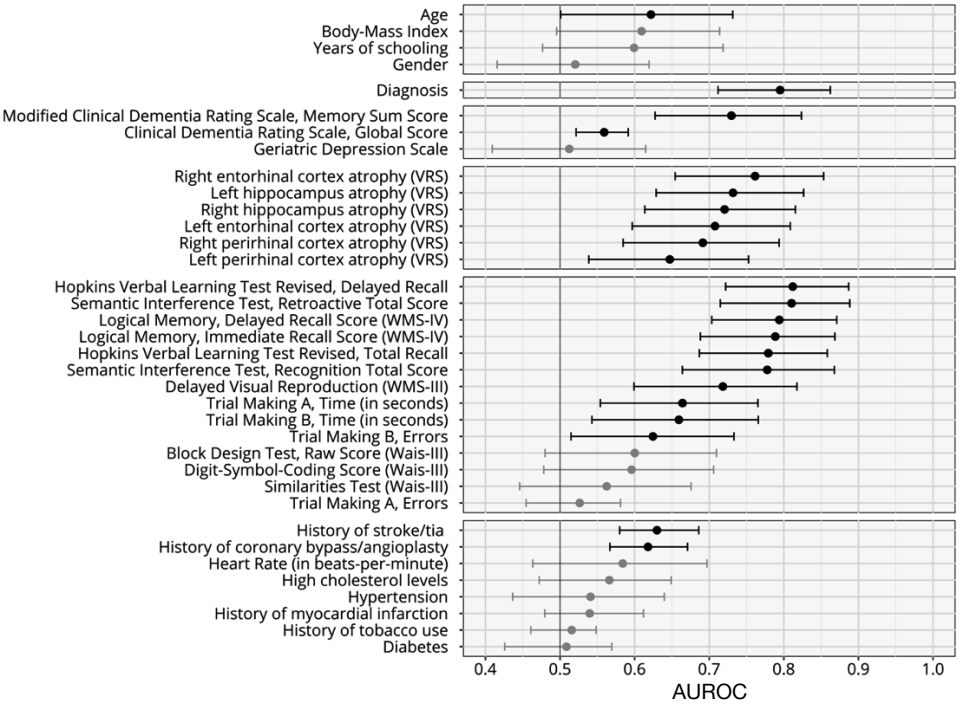
AUROC = Area Under the Receiving Operating Curve; SVM = Support Vector Machine.

Figure 1. AUROC of algorithms.



The figure indicates the cross-validated AUROC and its 95% bootstrap C.I. for each algorithm. Algorithms are grouped according to the machine learning techniques. The different feature selection procedure applied are indicated below, as well as by different point shapes (circle = all features; square = features selected via EN, triangle = features selected via RFE)

Figure 2. AUROC of individual predictors



The figure indicates the cross-validated AUROC and its 95% bootstrap C.I. when prediction is made by each single predictor. Predictors are grouped according to conceptual domains, in descending order sociodemographic information, diagnosis, clinical scores, brain atrophy, cognitive measures and cardiovascular risk index. Non-significant AUROC (i.e., lower bound of the C.I. lower than or equal to 0.5) are in grey, significant ones in black.

CHAPTER 3

A CLINICALLY-TRANSLATABLE MACHINE LEARNING ALGORITHM FOR THE PREDICTION OF ALZHEIMER'S DISEASE CONVERSION: FURTHER EVIDENCE OF ITS ACCURACY VIA A TRANSFER LEARNING APPROACH

Massimiliano Grassi[1], David A. Loewenstein[2,3,4], Daniela Caldirola[1], Koen Schruers[5], Ranjan Duara[2,3,7], Giampaolo Perna[1,2,5,8]

1 Department of Clinical Neurosciences, Hermanas Hospitalarias, Villa San Benedetto Menni Hospital, Albese con Cassano, Como, Italy.
2 Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, FL, USA.
3 Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami, FL, USA.
4 Center on Aging, Miller School of Medicine, University of Miami, FL, USA.
5 Research Institute of Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine and Life Sciences, University of Maastricht, Maastricht, Netherlands.
6 Department of Neurology, Herbert Wertheim College of Medicine, Florida International University, Miami, FL, USA.
7 Courtesy Professor of Neurology, Department of Neurology, University of Florida College of Medicine, Gainesville Florida
8 Mantovani Foundation, Arconate, Italy.

Reference: Grassi M, Loewenstein DA, Caldirola D, Schruers K, Duara R, Perna G. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach. Int Psychogeriatr. 2018 Nov 14:1-9.

Abstract

Background: in a previous study, we developed a highly performant and clinically-translatable machine learning algorithm for a prediction of 3-year conversion to Alzheimer's disease in subjects with Mild Cognitive Impairment (MCI) and Premild Cognitive Impairment. Further tests are necessary to demonstrate its accuracy when applied to subjects not used in the original training process. In this study, we aimed to provide preliminary evidence of this via a transfer learning approach.
Methods: we initially employed the same baseline information (i.e., clinical and neuropsychological test scores, cardiovascular risk indexes, and a visual rating scale for brain atrophy) and the same machine learning technique (support vector machine with radial-basis function kernel) used in our previous study to retrain the algorithm to discriminate between participants with AD (n=75) and normal cognition (n=197). Then, the algorithm was applied to perform the original task of predicting the 3-year conversion to AD in the sample of 61 MCI subjects that we used in the previous study.
Results: even after the retraining, the algorithm demonstrated a significant predictive performance in the MCI sample (AUROC = 0.821, 95% CI bootstrap = 0.705-0.912, best balanced accuracy = 0.779, sensitivity = 0.852, specificity = 0.706).
Conclusions: these results provide first indirect evidence that our original algorithm can perform relevant generalized predictions also when applied to new MCI individuals. This motivates future efforts to bring the algorithm at sufficient levels of optimization and trustworthiness that will allow its application in clinical as well as research settings.

Keywords: Alzheimer's disease, clinical prediction rule, machine learning, mild cognitive impairment, personalized medicine, precision medicine, transfer learning

Introduction

Alzheimer's disease (AD) is the most common neurodegenerative brain disorder and is the top cause for disabilities in later life being associated with huge global costs. Currently, no cure is available for AD although, with new emerging treatment approaches, it is increasingly important to be able to identify subjects at a true risk of later developing AD. By identifying those persons at greatest risk for decline, it would be possible to make clinical trials of AD treatments more cost-effective and valid by better selecting subjects to recruit, as treatments will likely have the greatest impact when provided at the earliest possible stage of the disease process (Brooks and Loewenstein 2010, Loewenstein, Curiel et al. 2017).

In a previous study, we presented a novel machine learning (ML) algorithm for the prediction of a three-year conversion to AD in subjects with Mild Cognitive Impairment (MCI) and preliminarily also in subjects with Premild Cognitive Impairment (PreMCI), which was developed with a sample of 123 MCI and PreMCI patients recruited in a collaborative longitudinal study by several centers located in the Miami (Florida, US) area (Grassi, Perna et al. 2018). Differently from several other available ML algorithms, ours employed only non-invasive predictors that are either already routinely assessed or effectively introducible in current clinical practice, i.e., clinical and neuropsychological test scores, cardiovascular risk indexes, and clinician-rated levels of brain atrophy. Promisingly, the algorithm achieved high predictive accuracy in our previous study, with a cross-validated balanced accuracy of 0.913 and Area Under the Receiving Operating Curve (AUROC) of 0.962 in the entire sample of MCI and PreMCI, and with a cross-validated balanced accuracy of 0.874 and AUROC of 0.914 in the sole sample of MCI. Its level of accuracy is among the very best of the many algorithms available in literature and the best achieved so far using only information easily collectable in clinical practice (Plant, Teipel et al. 2010, Apostolova, Hwang et al. 2014, Clark, Kapur et al. 2014, Agarwal, Ghanty et al. 2015, Moradi, Pepe et al. 2015, Dukart, Sambataro et al. 2016, Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017, Mathotaarachchi, Pascoal et al. 2017, Minhas, Khanum et al. 2017).

However, before an application can be safely proposed, a predictive algorithm needs to be tested in further independent samples of MCI and PreMCI subjects to demonstrate that its accuracy levels are preserved also when it is applied in generalized clinical and experimental contexts. To provide such evidence, a sample of distinct MCI and PreMCI subjects is currently under recruitment as part of a longitudinal study of over 450 persons at the University of Miami (DL). This sample will be used to test the algorithm we proposed in our previous study as soon as the three-year follow-up assessments will be completed.

However, before such optimal testing strategy will become performable, another opportunity for a preliminary test of our algorithm can come from the application of the so-called transfer learning approach. In the ML field, this refers to the process of using knowledge from one problem to train part or an entire algorithm that will be later applied to another problem. Its application for the solution of many complex tasks has been growing in the last years (Weiss, Khoshgoftaar et al. 2016) and such strategy has already been applied in developing several algorithms that predict the conversion of MCI subjects to AD (Nho, Shen et al. 2010, Plant, Teipel et al. 2010, Cui, Liu et al. 2011, Hinrichs, Singh et al. 2011, Cheng, Zhang et al. 2013, Westman, Aguilar et al. 2013, Young, Modat et al. 2013, Retico, Bosco et al. 2015, Collij, Heeman et al. 2016, Dukart, Sambataro et al. 2016). In these studies, the Authors initially trained the algorithms to discriminate between AD and Cognitively Normal individuals (CN), not using samples of MCI subjects but instead samples of solely AD and CN subjects. Then, they applied these ML algorithms in a different task, which is the prediction of the risk of future conversion to AD in MCI subjects. A prediction of conversion is made if the MCI subjects is classified as AD by the algorithm, while a prediction of not conversion is made if the MCI subjects is classified as CN. Such strategy is motivated by the hypothesis that those MCI subjects who will later convert to AD already show AD-like unrecognized characteristics, which instead do not characterize the MCI subjects who will not convert.

Following the abovementioned approach, we employed the same predictors and ML technique (support vector machine with radial-basis function kernel) used in our previous study to retrain our algorithm to discriminate between AD and CN using a sample of subjects with either the former or the latter condition. Then, after retraining, we will use our

ML algorithm to make a prediction of the three-years conversion to AD in the same sample of MCI subjects we used in our previous study. In the current study, although we will use the same predictors and ML technique, the MCI sample will be used only to test the algorithm and not during training. Thus, the results we will achieve will be able to provide first indirect evidence of how our previously proposed ML algorithm can perform relevant predictions also when applied to a sample of subjects not used in the training process.
Compared to our previous study where both training and validation were performed in the MCI sample via a cross-validation procedure, we expect that the algorithm retrained in a separate sample of AD/CN individuals will achieve a reduced but still relevant predictive performance in the MCI sample, which will provide further complementary evidence in support of the results found in our previous study.

Materials and methods

Subjects

Data regarding 272 subjects with AD or CN as well as the sample of 61 subjects with MCI used in our previous study (Grassi, Perna et al. 2018) were included in the current one. Instead, considering that only three converters were available in the PreMCI sample, this was employed in the current study. All the included sample of subjects are part of a dataset that collects several patients recruited in a study investigating longitudinal changes associated with MCI and normal aging, which involved community volunteers as well as subjects recruited from the Memory Disorders Clinic at the Wien Center for Alzheimer's disease, the Memory Disorders at Mount Sinai Medical Center, Miami Beach, Florida, and the community and memory disorders center at the University of South Florida. A common clinical and neuropsychological battery was administered to all subjects at all the sites. Considering the final aim of developing a predictive algorithm to be used in clinical practice, no other inclusion or exclusion criteria were applied beyond these diagnostic

criteria and the occurrence of missing information in the variables used as predictors (see below).
Subjects were classified as having probable Alzheimer's disease (AD; n = 75, 27.07%) if at the time of the assessment they presented a Dementia syndrome by DSM-IV-TR criteria (American Psychiatric Association 2000), and satisfied the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association criteria for Alzheimer's disease (McKhann, Drachman et al. 1984). Subjects were classified as MCI if they presented subjective memory complaints by the participant and/or or collateral informant, and evidence of decline from clinical history and evaluation, i.e., a global CDR score (Morris 1993) of 0.5, and one or more memory measures (including the HVLTR, the SIT, Logical Memory Delay and Visual Reproduction of the WMS-IV, TMT-B, Category Fluency, Letter Fluency and WAIS-III Block Design) 1.5 standard deviation or greater below expected normative values. Finally, subjects were identified as CN (n = 197, 72.03%) if during assessment they had a global CDR of 0 and no neuropsychological deficits (1.5 standard deviation or greater above expected normative values).
The study was conducted with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All subjects gave their written informed consent to the use of their clinical data for scientific research purposes.

Feature extraction

The same variables included as predictors in the best algorithm developed in our previous study were used to train the AD/CN algorithm, excluding the variable indicating the MCI/PreMCI sub-type that is not applicable to AD and CN subjects. These predictors were selected with a recursive feature elimination procedure starting from a larger set of 36 variables regarding sociodemographic characteristics, clinical and neuropsychological test scores, cardiovascular risk indexes, and a visual rating scale for brain atrophy.

- *Clinical scales:* the memory sum score of a modified informant-based version of CDR (ModCDR-M) (Duara, Loewenstein et al. 2010);
- *Visual Rating Scale for brain atrophy:* left and right hemisphere HPC, ERC, and PRC atrophy levels were assessed with a 0-4 VRS (Duara, Loewenstein et al. 2008), an adaptation from the original Scheltens' VRS for the global assessment of medial temporal atrophy (Scheltens, Leys et al. 1992). Ratings were performed on a Magnetic Resonance Imaging (MRI) image of a standardized coronal slice, perpendicular to the line joining the anterior and posterior commissures, intersecting the mammillary bodies and on adjacent slices. Ratings are performed on a five-point scale (0 = no atrophy, 1 = minimal atrophy, 2 = mild atrophy, 3 = moderate atrophy, and 4 = severe atrophy) and excellent inter-rater (kappa, 0.75 to 0.94) and intra-rater (kappa, 0.84 to 0.94) agreements have been reported (Duara, Loewenstein et al. 2008, Urs, Potter et al. 2009), also thanks to a computer interface that provides a library of reference images. Only five of the six VRS scores were included in the algorithm, excluding the left PRC score as indicated by the feature selection procedure used in our previous study.
- *Neuropsychological tests:* The Hopkins Verbal Learning Test Revised - Total Recall (HVLTR-R) and Delayed Recall (HVLTR-D) scores (Benedict and Zgaljardic 1998), the Semantic Interference Test - Total Retroactive (SIT-RT) and Total Recognition (SIT-RC) scores (Loewenstein, Acevedo et al. 2004), the Trial Making the Logical Memory Test - Immediate Recall (LM-I) and - Delayed Recall (LM-D) scores of the WMS-IV (Wechsler 1997).
- *Cardiovascular risk indexes:* heart rate, and history of myocardial infarction.

Detailed descriptions of these variables can be found in (Grassi, Perna et al. 2018). Continuous variables were standardized. In the end, 14 continuous and one dichotomous categorical predictor were used. The full list is available in Table 1.

Training with AD/CN Participants

In this study, we used the same ML technique that generated the most performing algorithm in our previous study (Grassi, Perna et al. 2018), which is Support Vector Machine with radial basis function (Gaussian) kernel (SVM). It has two hyper-parameters (σ; C) that allow a different tuning of the algorithm during the training process, and 200 random configurations of these hyper-parameters were attempted in order to identify the configuration that allow to achieve the best predictive performance.
Specifically, we are interested in achieving the hyper-parameter configuration that results in the best possible performance when the algorithm is applied to discriminate new AD/CN cases that are not part of the training sample. We used cross-validation to provide an estimate of such generalized performance but the sample size in this study resulted too large to apply the computationally expensive leave-pair-out cross-validation protocol, as we did in our previous study. Instead, a stratified cross-validation protocol was used. For each hyper-parameter configurations, 75 folds were used, each including a single subject with AD. Training was performed excluding the cases in the fold from the training sample and calculating the performance of the algorithm on them. Finally, the average performance metric is taken as estimate of the generalized performance of the algorithm created with that hyper-parameter configuration. As primary performance metric, the Area Under the Receiving Operating Curve (AUROC) was used. At first the algorithm outputs a continuous prediction score (range: 0-1; the closer to 1 the higher the predicted risk of conversion for that subject) and then the class prediction is finally made setting a cut-off score (AD if above or equal to the cut-off score, CN if below).
A bootstrap procedure, (10000 resampling with replacement) was used to calculate the confidence interval (CI) of the average cross-validated AUROC. The distribution of the resampled 10000 average AUROCs was used to calculate 95% CI with the bias-corrected and accelerated (BCa) approach (Efron 1987).
The hyper-parameter configuration for each technique that produced the best cross-validated AUROC was retained and a final algorithm with

such configuration is finally trained on the whole dataset of AD and CN subjects.

Testing with MCI

Predictions of three-year conversion to AD for the MCI subjects was obtained using the algorithm trained with AD and CN subjects, considering a classification of AD as prediction of future conversion to AD and CN as prediction of non-conversion. It is worth noting that in this case the MCI subjects were not used during the training of the model. The AUROC in the MCI sample subsample were calculated and a stratified bootstrap procedure, (10000 resampling with replacement) was used to calculate the AUROC confidence interval (CI). The distribution of the new 10000 AUROCs calculated was used to calculate 95% CI with the bias-corrected and accelerated (BCa) approach (Efron 1987). The cut-off applied to the algorithm output scores was progressively increased starting from 0 and the thresholds providing the best balanced accuracy was identified, calculating also the sensitivity and specificity achieved. Moreover, the cross-validated levels of specificities and balanced accuracy values when sensitivity approached to 0.95, 0.9, 0.85, 0.8, 0.75 were calculated.

Results

Descriptive statistics of each feature in the AD and CN groups are reported in Table 1. Statistics of continuous features are reported before standardization was applied.
The final algorithm trained with the AD/CN sample shows very high cross-validated accuracy in discriminating between AD and CN individuals, with an AUROC of 0.996 (C.I. 95% bootstrap= 0.983, 1).
When applied to sample of MCI individuals to predict their risk of conversion to AD in the next 3-year, its predictive performance resulted relevant also in this task, with an AUROC of 0.821 (C.I. 95% bootstrap = 0.705, 0.912) and a best balanced accuracy of 0.779 (sensitivity = 0.852, specificity = 0.706). The levels of specificities and balanced accuracy

values when sensitivity approached 0.95, 0.9, 0.85, 0.8, 0.75 are reported in Table 2.
As expected, its predictive performance was smaller than the cross-validated one found in our previous study (AUROC = 0.914, C.I. 95% bootstrap = 0.822, 0.975; best balanced accuracy = 0.874, sensitivity = 0.880, specificity = 0.867) but it demonstrated a predictive performance better than randomness (i.e., the AUROC has a lower 95% bootstrap CI larger than 0.5).

Discussion

The aim of the current study was to provide a first indirect evidence in support of a clinically-translatable machine-learning algorithm for the identification of three-year conversion to Alzheimer's disease in subjects with either MCI or PreMCI, which we presented in a previous paper (Grassi, Perna et al. 2018). Such algorithm showed high cross-validated predictive performance, the highest among the currently available algorithms that are based only on information easily assessable in clinical practice (Grassi, Perna et al. 2018).
A three-year follow-up assessment of a new sample of MCI subjects is currently ongoing, which will allow a proper testing of our algorithm in a sample that is independent from the one employed in the training phase. Instead, in this study, we used the transfer learning approach to preliminary perform such testing. We employed the same feature and ML techniques used in our previous study to retrain the algorithm to discriminate between AD and CN participants, and then we applied it to the sample of MCI subjects that we used in the previous study, considering a prediction of a three-year conversion to AD if the algorithm classifies a MCI subject as AD and a prediction of non-conversion if the algorithm classifies the subject as CN.
As hypothesized, after the retraining, the algorithm demonstrated a significant predictive performance in the MCI sample, although reduced in magnitude compared to the one achieved in our previous study (Grassi, Perna et al. 2018). These results suggest that our algorithm can perform relevant predictions also when applied to new samples not used

for training, further motivating future efforts to bring our algorithm at a clinical-ready level.

Previously, other investigators have applied a similar strategy to develop predictive algorithms for the conversion to AD in MCI subjects (Nho, Shen et al. 2010, Plant, Teipel et al. 2010, Cui, Liu et al. 2011, Hinrichs, Singh et al. 2011, Cheng, Zhang et al. 2013, Westman, Aguilar et al. 2013, Young, Modat et al. 2013, Retico, Bosco et al. 2015, Collij, Heeman et al. 2016, Dukart, Sambataro et al. 2016), based on the hypothesis that the MCI subjects who will later convert to AD already show characteristics of the AD, and that their MCI condition is caused by the same pathophysiological process that will later lead to a full-blow AD manifestation, which has already begun although not fully evident yet. Instead, the non-converters have MCI for other causes and their traits are distinct from those characterizing subjects with AD.

Results from previous studies that trained algorithms with an AD/CN sample and then applied them to predict the conversion to AD of MCI subjects are summarized in Table 3. Our algorithm achieved one of the best predictive performance available, with only the algorithms presented by Young and colleagues (Young, 2013 #75), and Dukart and colleagues (Dukart, 2016 #49) showing respectively the former similar, and the latter higher performances compared to ours. However, both these ML algorithms necessitate information that are currently not easily and routinely assessed in clinical practice, i.e., 18-fluorodeoxyglucose Positron Emission Tomography, and the typization of APOE gene. These results are consistent with the evidence from the previous study that our proposed algorithm is the best performing one among those based on only information easy to be clinically collected.

A reduced predictive performance compared to when the algorithm was trained directly on MCI and PreMCI subjects was expected. First, the training and the tuning of the model hyper-parameters were performed to accomplish a different classification task, i.e., distinguishing AD and CN subjects. Even if it is hypothesized that MCI converters show AD-like characteristics while non-converters do not, the AD/CN and converters/non-converters classification tasks may share a common but not totally equal solution. Thus, the optimized hyper-parameter configuration of a ML algorithm identified to perform the former may be a good but not the very best possible to perform the latter, taking to a

sub-optimal predictive accuracy. Moreover, training the algorithm with AD and CN subjects did not enable to include one of the predictors we have previously used, i.e., the MCI/PreMCI sub-type, which resulted of particular relevance for the prediction. The lack of this piece of information may also have caused part of the fall in the predictive performance compared to what previously achieved. Despite these abovementioned issues, the retrained algorithm achieved a significant predictive capability in the MCI sample, which in this study was not directly employed in the training phase.

However, it is worth highlighting that our MCI sample cannot be viewed as perfectly independent from the AD/CN training sample, which is a potential limitation of the current study. As a matter of facts, both samples have been recruited in the same clinical centers as part of the same longitudinal study, all located in the area of Miami. The population referring to this study might have peculiar characteristics and the performance of the algorithm might result partially reduced if used in different populations. Moreover, in the previous study, both the feature selection and the identification of the best ML technique was performed with a sample that also included the same MCI sample here applied as a test dataset. As in this study we used the same features and ML technique that were selected in our previous study, some minor so-called data leakage may indeed have occurred. In the ML field, this indicates that some information may have passed from the training to the test process, which can cause a partial inflation of the estimate of the algorithm generalized performance obtained by its application to the test set. The inflation is expected to be the more severe the greater the amount of information shared between training and testing, which in our analyses we expect to be limited and only related to the issues we have just discussed.

Albeit taking into account these limitations, the results of this study further support that the baseline information we took into account together with the use of ML techniques can effectively allow a prediction of conversion to AD in MCI subjects and they motivate to proceed in a further development and testing of the algorithm in order to reach sufficient levels of optimization and trustworthy for its application in clinical as well as research settings.

Specifically, some main issues will be principally addressed in the next phase: first, a test of the generalizability of the algorithm will be performed by applying it to new MCI subjects, which are currently under recruitment in a longitudinal study of over 450 persons at the University of Miami (DL). In addition, we aim to test the algorithm predictive performance in further subjects with PreMCI that convert to AD within three-years. This would allow the use of the algorithm to identify fast converters to AD at a very early stage of the degenerative process. Finally, a particular effort will be made to provide an explanation of which role each feature plays in the prediction. The current algorithm was chosen in our previous study because it proved to significantly outperform all the others we attempted. However, this algorithm results a “black-box” at the moment as it does not allow an easy interpretation of how the algorithm achieve to perform the predictions. A better interpretability of the algorithm will allow to foster its application by being better comprehended and accepted by all users, as well as it may allow to reach further potential insights regarding the development process of AD.

Acknowledgments: this study was supported by the National Institute on Aging, United States Department of Health and Human Services (grant: 1P50AG025711-05, R01AG047649-01A1, P50 AG047726602).

References

Agarwal, S., P. Ghanty and N. R. Pal (2015). "Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease." Bioinformatics **31**(15): 2505-2513.

American Psychiatric Association (2000). Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision). 4 ed. Washington, DC, Amer Psychiatric Pub.

Apostolova, L. G., K. S. Hwang, O. Kohannim, D. Avila, D. Elashoff, C. R. Jack, Jr., L. Shaw, J. Q. Trojanowski, M. W. Weiner, P. M. Thompson and I. Alzheimer's Disease Neuroimaging (2014). "ApoE4

effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer's disease." Neuroimage Clin **4**: 461-472.

Benedict, R. H. and D. J. Zgaljardic (1998). "Practice effects during repeated administrations of memory tests with and without alternate forms." J Clin Exp Neuropsychol **20**(3): 339-352.

Brooks, L. G. and D. A. Loewenstein (2010). "Assessing the progression of mild cognitive impairment to Alzheimer's disease: current trends and future directions." Alzheimers Res Ther **2**(5): 28.

Cheng, B., D. Zhang, S. Chen, D. I. Kaufer, D. Shen and I. Alzheimer's Disease Neuroimaging (2013). "Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers." Neuroinformatics **11**(3): 339-353.

Clark, D. G., P. Kapur, D. S. Geldmacher, J. C. Brockington, L. Harrell, T. P. DeRamus, P. D. Blanton, K. Lokken, A. P. Nicholas and D. C. Marson (2014). "Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease." Cortex **55**: 202-218.

Collij, L. E., F. Heeman, J. P. Kuijer, R. Ossenkoppele, M. R. Benedictus, C. Moller, S. C. Verfaillie, E. J. Sanz-Arigita, B. N. van Berckel, W. M. van der Flier, P. Scheltens, F. Barkhof and A. M. Wink (2016). "Application of Machine Learning to Arterial Spin Labeling in Mild Cognitive Impairment and Alzheimer Disease." Radiology **281**(3): 865-875.

Cui, Y., B. Liu, S. Luo, X. Zhen, M. Fan, T. Liu, W. Zhu, M. Park, T. Jiang, J. S. Jin and I. Alzheimer's Disease Neuroimaging (2011). "Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors." PLoS One **6**(7): e21896.

Duara, R., D. A. Loewenstein, M. T. Greig-Custo, A. Raj, W. Barker, E. Potter, E. Schofield, B. Small, J. Schinka, Y. Wu and H. Potter (2010). "Diagnosis and staging of mild cognitive impairment, using a modification of the clinical dementia rating scale: the mCDR." Int J Geriatr Psychiatry **25**(3): 282-289.

Duara, R., D. A. Loewenstein, E. Potter, J. Appel, M. T. Greig, R. Urs, Q. Shen, A. Raj, B. Small, W. Barker, E. Schofield, Y. Wu and H. Potter

(2008). "Medial temporal lobe atrophy on MRI scans and the diagnosis of Alzheimer disease." Neurology **71**(24): 1986-1992.

Dukart, J., F. Sambataro and A. Bertolino (2016). "Accurate Prediction of Conversion to Alzheimer's Disease using Imaging, Genetic, and Neuropsychological Biomarkers." J. Alzheimers. Dis. **49**(4): 1143-1159.

Efron, B. (1987). "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association **82**(397): 171-185.

Grassi, M., G. Perna, D. Caldirola, K. Schruers, R. Duara and D. A. Loewenstein (2018). "A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment." Journal of Alzheimer's Disease **61**(4): 1555-1573.

Hinrichs, C., V. Singh, G. Xu, S. C. Johnson and I. Alzheimers Disease Neuroimaging (2011). "Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population." Neuroimage **55**(2): 574-589.

Hojjati, S. H., A. Ebrahimzadeh, A. Khazaee, A. Babajani-Feremi and I. Alzheimer's Disease Neuroimaging (2017). "Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM." J. Neurosci. Methods **282**: 69-80.

Loewenstein, D. A., A. Acevedo, C. Luis, T. Crum, W. W. Barker and R. Duara (2004). "Semantic interference deficits and the detection of mild Alzheimer's disease and mild cognitive impairment without dementia." J Int Neuropsychol Soc **10**(1): 91-100.

Loewenstein, D. A., R. E. Curiel, R. Duara and H. Buschke (2017). "Novel Cognitive Paradigms for the Detection of Memory Impairment in Preclinical Alzheimer's Disease." Assessment: 1073191117691608.

Long, X., L. Chen, C. Jiang, L. Zhang and I. Alzheimer's Disease Neuroimaging (2017). "Prediction and classification of Alzheimer disease based on quantification of MRI deformation." PLoS One **12**(3): e0173372.

Mathotaarachchi, S., T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, V. S. Fonov, S. Gauthier, P. Rosa-Neto and I. Alzheimer's Disease Neuroimaging (2017). "Identifying incipient

dementia individuals using machine learning and amyloid imaging." Neurobiol Aging **59**: 80-90.

McKhann, G., D. Drachman, M. Folstein, R. Katzman, D. Price and E. M. Stadlan (1984). "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease." Neurology **34**(7): 939-944.

Minhas, S., A. Khanum, F. Riaz, A. Alvi and S. A. Khan (2017). "A Nonparametric Approach for Mild Cognitive Impairment to AD Conversion Prediction: Results on Longitudinal Data." IEEE J Biomed Health Inform **21**(5): 1403-1410.

Moradi, E., A. Pepe, C. Gaser, H. Huttunen, J. Tohka and I. Alzheimer's Disease Neuroimaging (2015). "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects." Neuroimage **104**: 398-412.

Morris, J. C. (1993). "The Clinical Dementia Rating (CDR): current version and scoring rules." Neurology **43**(11): 2412-2414.

Nho, K., L. Shen, S. Kim, S. L. Risacher, J. D. West, T. Foroud, C. R. Jack, M. W. Weiner and A. J. Saykin (2010). "Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging." AMIA Annu Symp Proc **2010**: 542-546.

Plant, C., S. J. Teipel, A. Oswald, C. Bohm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel and M. Ewers (2010). "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease." Neuroimage **50**(1): 162-174.

Retico, A., P. Bosco, P. Cerello, E. Fiorina, A. Chincarini and M. E. Fantacci (2015). "Predictive Models Based on Support Vector Machines: Whole-Brain versus Regional Analysis of Structural MRI in the Alzheimer's Disease." J Neuroimaging **25**(4): 552-563.

Scheltens, P., D. Leys, F. Barkhof, D. Huglo, H. C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters and J. Valk (1992). "Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates." J Neurol Neurosurg Psychiatry **55**(10): 967-972.

Urs, R., E. Potter, W. Barker, J. Appel, D. A. Loewenstein, W. Zhao and R. Duara (2009). "Visual rating system for assessing magnetic resonance images: a tool in the diagnosis of mild cognitive impairment and Alzheimer disease." J Comput Assist Tomogr **33**(1): 73-78.

Wechsler, D. (1997). WMS-III: Wechsler memory scale administration and scoring manual, Psychological Corporation.

Weiss, K., T. M. Khoshgoftaar and D. Wang (2016). "A survey of transfer learning." Journal of Big Data **3**(1): 9.

Westman, E., C. Aguilar, J. S. Muehlboeck and A. Simmons (2013). "Regional magnetic resonance imaging measures for multivariate analysis in Alzheimer's disease and mild cognitive impairment." Brain Topogr **26**(1): 9-23.

Young, J., M. Modat, M. J. Cardoso, A. Mendelson, D. Cash, S. Ourselin and I. Alzheimer's Disease Neuroimaging (2013). "Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment." Neuroimage Clin **2**: 735-745.

Table 1. Descriptive statistics.

Continuous Predictors	Non-converters		Converters	
	Mean	S.D.	Mean	S.D.
Modified Clinical Dementia Rating Scale, Memory Sum Score	0.88	1.11	4.49	2.25
Right hippocampus atrophy (VRS)	0.56	0.68	1.79	1.23
Left hippocampus atrophy (VRS)	0.48	0.64	1.96	1.21
Right entorhinal cortex atrophy (VRS)	0.30	0.60	1.41	1.19
Left entorhinal cortex atrophy (VRS)	0.38	0.65	1.73	1.27
Right perirhinal cortex atrophy (VRS)	0.32	0.62	1.39	1.30
Hopkins Verbal Learning Test Revised, Total Recall	25.62	4.42	14.13	4.93
Hopkins Verbal Learning Test Revised, Delayed Recall	9.16	2.05	2.12	2.69
Semantic Interference Test, Retroactive Total Score	25.66	2.16	13.43	6.97
Semantic Interference Test, Recognition Total Score	28.40	1.74	18.55	6.74
Logical Memory, Immediate Recall Score (WMS-IV)	13.03	3.47	5.45	3.54
Logical Memory, Delayed Recall Score (WMS-IV)	11.38	3.42	2.96	3.22
Trial Making B, Errors	0.69	0.98	2.56	3.56
Heart Rate (in beats-per-minute)	69.72	8.40	70.37	9.44

Categorical Predictors		Non-converters		Converters	
		N	%	N	%
History of myocardial infarction	No	6	3.05%	7	9.33%
	Yes	191	96.95%	68	90.67%

S.D = Standard Deviation; N = number of subjects; aMCI = amnestic Mild Cognitive Impairment; non-aMCI = non-amnestic Mild Cognitive Impairment; PreMCI-cl = Premild Cognitive Impairment - clinical subtype; PreMCI-np = Premild Cognitive Impairment - neuropsychological subtype; WMS-IV = Weschler Memory Scale – Fourth Edition; VRS = Visual Rating Scale.

Table 2. Performance of the algorithm in the MCI sample.

	Performance Level (Target)	Sensitivity	Specificity	Balanced Accuracy
MCI (AUROC = 0.821)	Sensitivity of .95	0.963	0.471	0.717
	Sensitivity of .90	0.923	0.559	0.741
	Sensitivity of .85	0.852	0.706	0.779
	Sensitivity of .80	0.815	0.735	0.775
	Sensitivity of .75	0.778	0.735	0.756
	Best Balanced Accuracy	0.852	0.706	0.779

AUROC = Area Under the Receiving Operating Curve, MCI = Mild Cognitive Impairment.

Table 3. Comparison with previous algorithms.

Reference	Follow-up period	Predictors	Machine Learning Technique	AUROC	Specificity	Sensitivity	Balanced accuracy
Our study	3 years	Socio-demographic, clinical, neuropsychological, clinician-rated brain atrophy, cardiovascular risk scores	SVM with radial-basis function kernel	0.821	0.852	0.706	0.779
(Collij, Heeman et al. 2016)	1-4 years	Arterial spin labeling perfusion MRI	Linear SVM	0.71	0.75	0.67	n.a.
(Nho, Shen et al. 2010)	3 years	Brain structural MRI	SVM with radial-basis function kernel	n.a.	0.688	0.753	n.a.
(Retico, Bosco et al. 2015)	2 years	Brain structural MRI	Linear SVM	0.707	n.a.	n.a.	n.a.
(Westman, Aguilar et al. 2013)	1.5 years	Brain structural MRI	Partial least square to latent structures multivariate analysis	0.749	0.645	0.77	n.a.
(Dukart, Sambataro et al. 2016)	at least 2 years	Brain structural MRI, 18-fluorodeoxyglucose Positron Emission Tomography, and APOE typization	Naïve Bayes	0.84	0.861	0.875	0.868

(Plant, Teipel et al. 2010)	Approximately 2.5 years	Brain structural MRI	Voting Feature Intervals	n.a.	0.87	0.56	n.a.
(Cheng, Zhang et al. 2013)	not specified	Brain structural MRI, 18-fluorodeoxyglucose Positron Emission Tomography, and cerebrospinal fluid markers	multimodal relevance vector regression	-	0.541	0.641	-
(Cui, Liu et al. 2011)	2 years	Brain structural MRI, cerebrospinal fluid markers, neuropsychological, and functional activity scores	SVM with radial-basis function kernel	0.796	0.482	0.964	-
(Young, Modat et al. 2013	3 years	Brain structural MRI, 18-fluorodeoxyglucose Positron Emission Tomography, and APOE gene typization	GP	0.823	0.571	0.833	0.722
(Hinrichs, Singh et al. 2011)	2-3 years	Brain structural MRI, and 18-fluorodeoxyglucose Positron Emission Tomography	Multi-kernel SVM	0.738	-	-	-

AUROC = Area Under the Receiving Operating Curve; SVM = Support Vector Machine. If several algorithms were presented in the paper, only the results of the best one is presented here.

CHAPTER 4

A NOVEL ENSEMBLE-BASED MACHINE LEARNING ALGORITHM TO PREDICT THE CONVERSION FROM MILD COGNITIVE IMPAIRMENT TO ALZHEIMER'S DISEASE USING SOCIO-DEMOGRAPHIC CHARACTERISTICS, CLINICAL INFORMATION AND NEUROPSYCHOLOGICAL MEASURES

Massimiliano Grassi[6,1,2], Nadine Rouleaux[§,3], Daniela Caldirola[1,2], David Loewenstein[4,5,6], Koen Schruers[7], Giampaolo Perna[2,1,4,7], Michel Dumontier[3], for the Alzheimer's Disease Neuroimaging Initiative[#]

1 Department of Clinical Neurosciences, Hermanas Hospitalarias, Villa San Benedetto Menni Hospital, Albese con Cassano, Como, Italy.
2 Department of Biomedical Sciences, Humanitas University, Rozzano, Milan, Italy.
3 Institute of Data Science, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands.
4 Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, FL, USA.
5 Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center, Miami, USA.
6 Center for Cognitive Neuroscience and Aging, Miller School of Medicine, University of Miami, FL, USA.
7 Research Institute of Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands.

Reference: Grassi M, Rouleaux N, Caldirola D, Loewenstein D, Schruers K, Perna G, Dumontier M; Alzheimer's Disease Neuroimaging Initiative. A Novel Ensemble-Based Machine Learning Algorithm to Predict the Conversion From Mild Cognitive Impairment to Alzheimer's Disease Using Socio-Demographic Characteristics, Clinical Information, and Neuropsychological Measures. Front Neurol. 2019 Jul 16;10:756.

---

6 Co-first authors

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/themes/freshnews-dev-v2/documents/policy/ADNI_Acknowledgement_List%205-29-18.pdf

Abstract

Background: Despite the increasing availability in brain health related data, clinically translatable methods to predict the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's disease (AD) are still lacking. Although MCI typically precedes AD, only a fraction of 20-40% of MCI individuals will progress to dementia within 3 years following the initial diagnosis. As currently available and emerging therapies likely have the greatest impact when provided at the earliest disease stage, the prompt identification of subjects at high risk for conversion to AD is of great importance in the fight against this disease. In this work, we propose a highly predictive machine learning algorithm, based only on non-invasively and easily in-the-clinic collectable predictors, to identify MCI subjects at risk for conversion to AD.
Methods: The algorithm was developed using the open dataset from the Alzheimer's Disease Neuroimaging Initiative (ADNI), employing a sample of 550 MCI subjects whose diagnostic follow-up is available for at least 3 years after the baseline assessment. A restricted set of information regarding sociodemographic and clinical characteristics, and neuropsychological test scores was used as predictors and several different supervised machine learning algorithms were developed and ensembled in a final algorithm. A site-independent stratified train/test split protocol was used to provide an estimate of the generalized performance of the algorithm.
Results: The final algorithm demonstrated an AUROC of 0.88, sensitivity of 77.7%, and a specificity of 79.9% on excluded test data. The specificity of the algorithm was 40.2% for 100% sensitivity.
Conclusions: The algorithm we developed achieved sound and high prognostic performance to predict AD conversion using easily clinically derived information that makes the algorithm easy to be translated into practice. This indicates beneficial application to improve recruitment in clinical trials and to more selectively prescribe new and newly emerging early interventions to high AD risk patients.

Keywords: Alzheimer's disease, clinical prediction rule, machine learning, mild cognitive impairment, personalized medicine, precision medicine, neuropsychological tests.

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease characterized by progressive memory loss, cognitive impairment, and general disability; AD is the most common cause of dementia of the Alzheimer's type. The progression of AD comprises a long, unnoticed preclinical stage, followed by a prodromal stage of Mild Cognitive Impairment (MCI) that leads to severe dementia and eventually death (Alzheimer's Disease 2018). While no disease-modifying treatment is currently available for AD, a large number of drugs are in development and encouraging early-stage results from clinical trials provide for the first time a concrete hope that one or more therapies may become available in a few years (Liu, Hlávka et al. 2017). As the progression of the neuropathology in AD starts years in advance before clinical symptoms of the disease become apparent and progressive neurodegeneration has irreversibly damaged the brain, emerging treatments will likely have the greatest effect when provided at the earliest disease stages. Thus, the prompt identification of subjects at high risk for conversion to AD is of great importance.

The ability to identify declining individuals at the prodromal AD stage provides a critical time window for early clinical management, treatment & care planning and design of clinical drug trials (Alzheimer's 2018). Precise identification and early treatment of at-risk subjects would stand to improve outcomes of clinical trials and reduce healthcare costs in clinical practice. However, simulations also suggest that the health care system is not prepared to handle the potentially high volume of patients who would be eligible for treatment (Liu, Hlávka et al. 2017).

MCI represents (currently) the earliest clinically detectable stage of a potential ongoing progression towards AD or other dementias. The cognitive decline in MCI is abnormal given an individual's age and education level, but does not interfere with daily activities, and thus does not meet criteria for AD. However, only 20-40% of individuals will progress to AD within three years, with a lower rate of conversion reported in epidemiologic samples than in clinical ones (Petersen, Parisi et al. 2006, Roberts, Knopman et al. 2014).

Currently, there are no means to provide patients diagnosed with MCI with an early prognosis for conversion to AD. While changes in several biomarkers prior to developing AD have been reported, no single

biomarker appears to adequately predict the conversion from MCI to AD with an acceptable level of accuracy. As such, there is increasing evidence that the use of a combination of biomarkers can best predict the conversion to AD (Devanand, Liu et al. 2008, Sperling and Johnson 2013, Dukart, Sambataro et al. 2016, Giannakopoulos 2017, Alzheimer's 2018).

In the current age of big data and artificial intelligence technologies, considerable effort has been dedicated in developing machine learning algorithms that can predict the conversion to AD in subjects with MCI. In almost all medical fields, the introduction into research and clinical practice of machine learning based decision-making tools, and more in general the shift towards a personalized medicine paradigm, is currently a debated topic and viewed as an opportunity to improve clinical outcomes. Such objective tools may provide individual predictions with a certain degree of confidence based on information that can be collected about the subject, so that researchers and clinicians may be supported by these predictions in order to take better and more effective decisions (Perna, Grassi et al. 2018).

So far, many studies focused on predicting the conversion of AD in MCI patients using different combinations of data including brain imaging, CSF biomarkers, genotyping, demographic and clinical information, and cognitive performance, achieving varying levels of accuracy (Plant, Teipel et al. 2010, Apostolova, Hwang et al. 2014, Clark, Kapur et al. 2014, Agarwal, Ghanty et al. 2015, Moradi, Pepe et al. 2015, Clark, McLaughlin et al. 2016, Dukart, Sambataro et al. 2016, Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017, Minhas, Khanum et al. 2017; see Grassi, Loewenstein et al. 2018, Grassi, Perna et al. 2018 for a recent review of the most performing algorithms presented in the scientific literature so far). However, while combining different biomarkers improves model accuracy, there is a lack of consistency regarding a specific combined AD prediction model and a translation into practice is still lacking. One possible reason for this is that current algorithms generally rely on expensive and/or invasive predictors, such as brain imaging or CSF biomarkers. As such, these studies only serve the purpose of a proof-of-concept, without being further tested in independent and clinical samples.

The current study aimed to develop a clinically translatable machine learning algorithm to predict the conversion to AD in subjects with MCI within a 3-year period, based on fast, easy, and cost-effective predictors. Specifically, we chose to develop a variety of machine learning algorithms based on distinct supervised machine learning techniques and subsets of the considered predictors, followed by a weighted average rank ensemble strategy on the predictions provided by the various algorithms to obtain a final, more accurate prediction. Our hypothesis was that high predictive accuracy could be obtained using the above-mentioned approach with simple and non-invasive predictors. We used data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI; http://adni.loni.usc.edu/) with a particular consideration for socio-demographic and clinical information, and neuropsychological test scores rather than using complex, invasive, and expensive imaging or CSF predictors.

Materials and methods

ADNI

Data used in the preparation of this article were obtained from the ADNI database. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. It contains data of a large number of cognitive normal, MCI, and AD subjects recruited in over 50 different centers in US and Canada with follow-up assessments performed every 6 months.
For this study, we used a subset of the ADNI dataset called ADNIMERGE that includes a reduced selection of more commonly used variables (i.e., demographic, clinical exam total scores, MRI and PET variables). This subset is part of the official dataset provided by ADNI.

Subjects

Data regarding 550 subjects with MCI and with available diagnostic follow-up assessments for at least three years were included in the study. The most relevant inclusion criteria of ADNI studies are the following: age between 55-90; six grade education or work history; subjects had to be fluent English/Spanish speakers; Geriatric Depression Scale score less than 6; good general health; no use of excluded medications (e.g., medications with anticholinergic properties) and stability for at least 4 weeks of other allowed medications; Hachinski ischemic score scale less than or equal to 4. A complete description of the ADNI study inclusion/exclusion criteria, including the full list of excluded and permitted medications, can be found in the ADNI General Procedure Manual, pages 20-25 (link: https://adni.loni.usc.edu/wp-content/uploads/2010/09/ADNI_GeneralProceduresManual.pdf).

The diagnosis of MCI was performed with the following criteria: memory complaint by subject or study partner that is verified by a study partner; abnormal memory function documented by scoring below the education adjusted cutoff on the Logical Memory II subscale (Delayed Paragraph Recall) from the Wechsler Memory Scale – Revised, which is less than or equal to 11 for 16 or more years of education, less than or equal to 9 for 8-15 years of education, and less than or equal to 6 for 0-7 years of education; Mini-Mental State Exam (MMSE) score between 24 and 30; Clinical Dementia Rating (CDR) score of 0.5; Memory Box score at least of 0.5; general cognition and functional performance sufficiently preserved such that a diagnosis of AD cannot be made.

Subjects were classified as converters to probable AD (cAD; n = 197, 35.82%) if they satisfied the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association criteria for AD [28] during at least one of the follow-up assessments occurred within three years from the baseline investigation, as well as having a MMSE score between 20 and 2. Otherwise, they were classified as non-converters to AD (NC; n = 353, 64.18%).

The study procedures were approved by the institutional review boards of all participating centers to the Alzheimer's Disease Neuroimaging

Initiative, and written informed consent was obtained from all participants or their authorized representatives.

Feature extraction

Considering our aim to employ only predictors that are either already routinely assessed or easily introducible in clinical practice, and that are not perceived as invasive by patients, we decided to take into account only variables in the ADNIMERGE dataset that regards diagnostic subtypes, sociodemographic characteristics, clinical and neuropsychological test scores. Some of these variables were not available for all recruited subjects and it was a priori decided to remove variables with greater than 20% missing values. Only the Digit Span Test score (DIGIT) exceeded the cut-off (52.73%) and was not used in our analysis. The following variables were used:

- Sociodemographic characteristics: sex, age (in years), years of education, and marital status (never married, married, divorced, widowed, unknown).\
- Subtypes of MCI: Early or Late MCI according to their score in the Logic Memory subscale of the Wechsler Memory Scale - Revised (Wechsler 1997), adjusted for the years of education. 9-11 Early MCI and ≤8 Late MCI for 16 or more years of education; 5-9 Early MCI and ≤4 Late MCI for 8-15 years of education; 3-6 Early MCI and ≤2 Late MCI for 0-7 years of education.
- Clinical scales: CDR (Morris 1993) was used to characterize six domains of cognitive and functional performance in AD and related dementias: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. The rating is obtained through a semi-structured interview of the patient together with other informants (e.g., family members). Sum of Boxes score was used in the current analyses (CDRSB). The score of the Functional Assessment Questionnaire (FAQ) (Pfeffer, Kurosaki et al. 1982), an informant-based clinician-administered questionnaire which assess the functional daily-living impairment in dementia, was also used in the analyses.

- Neuropsychological tests: MMSE (Folstein, Folstein et al. 1975) is a 30-point questionnaire that is used measuring cognitive impairment. All MCI subjects has a score of 24 of more at baseline. The Cognitive Subscale Alzheimer's Disease Assessment Scale (ADAS) (Rosen, Mohs et al. 1984) is made of 11 tasks that include both subject-completed tests and observer-based assessments, assessing the memory, language, and praxis domains. The result is a global final score ranging from 0 to 70, based on the sum of the scores of the single tasks (ADAS11). Beyond the ADAS11 score, the ADNI study also included an additional test of delayed word recall and a number cancellation or maze task, which are further summed to have a new total score that ranges from 0 to 85 (ADAS13). In addition, the score of the task 4 (Word Recognition, ADASQ4) was included in the ADNIMERGE dataset. All these three ADAS scores were initially considered as predictors in the analyses. The Rey Auditory Verbal Learning Test (RAVLT) (Schmidt and Others 1996) is a cognitive test used to evaluate verbal learning and memory. All the immediate (RAVLT-I), learning (RAVLT-L), forgetting (RAVLT-F), and percent forgetting (RAVLT-PF) scores were included in the ADNIMERGE dataset and used in the analyses. Moreover, the total delayed recall score of the Logic Memory subtest of the of the Wechsler Memory Scale-Revised (Wechsler 1945) (LDT), which assess verbal memory, and the time to complete of the Trial Making Test version B (TMTBT) (Reitan 1958), which assess visual-motor coordination and attentive functions. A summary of the abbreviations of all neuropsychological tests can be found in Table 1.

Taken together, 14 continuous, 2 dichotomous and 1 polytomous categorical features were initially considered. The full list is available in Table 2.

Dataset division in 5 site-independent, stratified test subsets

The entire dataset was divided in five mutually exclusive data subsets. These five subsets were created in order to satisfy the following criteria:

every subset has to include roughly 20% of the cases; all subjects from each of the 58 different recruitment sites has to be allocated into the same subset; every subset has to include roughly the same percentage of cAD as observed in the entire dataset (35.82%). In order to accomplish a division in 5 folds which satisfies all these criteria, 10000 different subsets were generated by progressively adding all subjects from a randomly chosen recruiting site, until the included cases ranged between 19% and 21% of the entire sample. Then, only those subsets whose percentage of cAD ranged between 35.52% and 36.12% were retained, which was satisfied in 567 (5,67%) out of the generated subsets. Finally, all possible combinations of five of the retained subsets were created in order to identify whether in any of these combinations covered the entire dataset without any repetition of cases. The entire process took around 4 hours of computation (on a Linux server with 2.20GHz Intel Xeon E5-2650 v4 CPUs), and successfully found a single combination of five subsets that satisfied all the desired criteria (Table 3).

All the missing value imputation, feature transformation and selection procedures, model training with cross-validation, and ensembling of different algorithms predictions described in the following paragraphs were performed in five distinct repetitions (named A-E) of the analyses, each time using the cases included in four of the five subsets and blindly to the remaining subset that were used as a test subset. The same missing value imputation, feature transformation and selection applied during training in the other four subsets were applied to the test subset. The predictive algorithms and their ensembling procedure developed in the other 4 subsets were tested against the test subset to obtain an estimate of the generalized performance in an independent sample of cases recruited in sites different from the ones used for training[7].

7 Our approach based on a division in 5 site-independent test subsets and a 10-fold cross-validation applied within each of them actually mimics the popular nested cross-validation approach, which is based on the nesting of an inner (in our case, the 10-fold cross-validation) and outer cross-validation loops (in our case, the site-independent test subsets). However, even if identical in its structure, in our study we did not compare the outer loop performances obtained re-applying nested cross-validation to different competing strategies (i.e., different machine learning techniques or ensembling approaches) in order to identify the best algorithmic approach, which is the primary reason for which any type of (cross-)validation strategy is employed. Instead, we a-priori chose to use all the 52 models we developed and to ensemble them with the average weighted ranks strategy. Thus, differently from what is done with nested cross-validation, the performance we observed in the outer loop was not used to take any choice about the development of the algorithm but only to provide a final estimate of the performance of our algorithm. For this reason, such final performance estimate can be safely considered a test instead of a validation of the performance of our algorithm, making the use of the term nested cross-validation not entirely appropriate and potentially misleading.

Feature transformation and selection

Imputation was performed for variables with missing values using the median for continuous features and using the mode for categorical features. Continuous variables were standardized (mean = 0, standard deviation = 1) and non-dichotomous categorical variables were dichotomized using one-hot encoding, i.e., re-coding them in a new dichotomous variable for each class of the categorical variable, with 1 indicating the occurrence of that class and 0 the occurrence of any other class of the variable.
In case groups of variables resulted highly correlated (pairwise r >= .75), principal component analysis was used to calculate principal components and the original variables were substituted with all the components with eigenvalues >= 1.
All features were initially used during training (feature set 1). Moreover, three feature subsets were additionally created based on different selection strategies in order to include only those that are the most informative. A filtering procedure was applied to create reduced sets of features based on their bivariate statistical association ($p < .05$) with the outcome using independent sample t-test for continuous predictors and Fisher's exact test for both dichotomous and one-hot encoded polytomous features (feature subset 2). Two cross-validated recursive feature elimination procedures (also known as "wrapper" procedures) with Logistic Regression (LR, feature subset 3) and Random Forest (RF, feature subset 4) (Breiman 2001) were also applied. In particular, the latter strategy was chosen because it has previously proved to be efficacious in selecting a relevant feature subset (Grassi, Perna et al. 2018).

Machine learning techniques

Several machine learning procedures that can be used to solve classification problems exists. We used 13 supervised techniques: LR, Naive Bayes (NB) (Rish and Others 2001), L1 and L2 regularized logistic regression or Elastic Net (EN) (Zou and Hastie 2005), Support Vector Machine (Schölkopf, Smola et al. 2002) with linear (SVM-Linear), radial basis function (SVM-RBF), and polynomial (SVM-Poly) kernels with Platt scaling (Platt 1999), k-Nearest Neighbors algorithm (kNN) (Altman 1992),

Multi-Layer Perceptrons with either one or two hidden layers and trained with either a full-batch gradient descent or adam (Kingma and Ba 2014) algorithms (MLP1-Batch, MLP2-Batch, MLP1-Adam, MLP2-Adam), RF, and Gradient boosted decision trees (GBDT) (Mason, Baxter et al. 2000). All analyses were parallelized on a Linux server equipped with four 12-core Intel Xeon CPU E5-2650 v4 @ 2.20GHz and were performed in Python 3.6 (Python Software), using the implementation of the machine learning techniques available in the Scikit-Learn library (Pedregosa, Varoquaux et al. 2011).

Hyper-parameter optimization

Machine learning techniques usually have one or more hyper-parameters that allow a different tuning of the algorithm during the training process. Different values of these hyper-parameters lead to algorithms with different predictive performances with the goal of obtaining the best possible performance when applied to cases that are not part of the training set. In order to optimize such hyper-parameters for each ML techniques used in this study, each model was trained with 50 random hyper-parameter configurations, and 50 further configurations were progressively estimated with a Bayesian optimization approach. Instead of a random generation, Bayesian optimization aims to estimate which is the hyper-parameter configuration that would maximize the performance of the algorithm starting from the previously attempted ones, based on the assumption that it exists a relationship between the various hyper-parameter values and the performance achieved by the algorithm. Bayesian optimization is expected of being able to identify better hyper-parameter configurations, and in a reduced number of attempts, than just trying to generate them at random. Estimation was performed with Gaussian Processes, as implemented in the Scikit-Optimized library (https://scikit-optimize.github.io/).
The Area Under the Receiving Operating Curve (AUROC) was used as performance metric to be maximized. All the ML algorithms developed in this study output a continuous prediction score (range: 0-1; the closer to 1 the higher the predicted risk of conversion for that subject) and the AUROC value can be interpreted as the probability that a randomly selected cAD subject will receive a higher output score than a randomly

selected NC subject. The AUROC value is 0.5 when the algorithm makes random predictions and 1 in case it is always correct in making predictions. AUROC is not affected by class imbalance, and it is independent with respect to any specific threshold that is applied to perform a dichotomous prediction.

Cross-validation procedure

The aim is to develop an algorithm that can achieve the best possible generalized performance and not to perform well only with the cases used in the training process. Cross-validation provides an estimate of such generalized performance for every hyper-parameter configuration. In cross-validation, the train sample is divided in several folds of cases that are held-out from the training process, with training iteratively performed with the remaining cases. After the training, the algorithm is finally applied on the held-out cases.
We applied the commonly used 10-fold cross-validation procedure, repeated 10 times to obtain a stable performance estimate. The fold creation was performed at random, stratifying (i.e., balancing) for the percentage of converters and non-converters in each fold. Finally, the 100 performance estimates of the algorithm available for each hyper-parameter configuration were averaged to provide a final point estimate of the generalized performance. The hyper-parameter configuration for each machine learning technique that demonstrated the best average cross-validated AUROC was retained.

Weighted rank average of single algorithm predictions

Using a collection of algorithms and combining their predictions instead of considering only the prediction coming from a single algorithm generally improves the overall predictive performance (Opitz and Maclin 1999). This procedure is called ensembling and it is also the principles on which some individual techniques such as Random Forest and Gradient Boosting techniques are based.
Several different ensemble methods exist, which usually require a further independent data subset from both the training and test ones. This additional subset would be used to train how to optimally combine the various predictions generated by the single algorithms. Given the limited

amount of data available in the current study, further reducing the size of the train sample may have undermined the predictive performance of the developed algorithms. Thus, we decided to apply a simple form of ensembling based on a weighted average of the rank predictions generated by all individual algorithms. This strategy is usually considered effective even though it does not require to develop any further machine learning meta-algorithm and to optimize its hyperparameters (Wolpert 1992).
First, the ranks of the cross-validated continuous prediction scores of the train subset cases were calculated for each of the 52 developed algorithms and rescaled in order to range between 0 and 1. Then, the arithmetic average of the rescaled ranks weighted for the cross-validated AUROC was calculated for each train subset case, representing the new continuous prediction scores for the train subset cases.
To generate the final continuous prediction scores of the test subset cases, at first 52 prediction scores for each test case were generated using all the 52 used algorithms. Then, the prediction score of each algorithm was substituted with the rescaled rank of the closest cross-validated train subset prediction score of that algorithm. Finally, the average of the rescaled ranks weighted for the cross-validated AUROC was calculated. This represents the final continuous prediction scores of each test subset cases.

Testing performance

The final continuous prediction scores of the five test subsets, which were obtained using the weighted rank average, were pooled and used to calculate the whole sample test AUROC. This represents the final estimate of the generalized site-independent AUROC that the algorithm is expected to achieve when it is applied to new cases. The 95% confidence interval (CI) of the AUROC was calculated with a stratified bootstrap procedure, with 10000 resamples and applying the bias-corrected and accelerated (BCa) approach (Efron 1987).
Different categorical cAD/NC predictions were generated for each case applying various thresholds to the final continuous prediction scores (i.e., a score equal or above the threshold indicated a cAD, otherwise a NC). First, the threshold values that maximized the balanced accuracy (i.e.,

the average between sensitivity and specificity) of the cross-validated train subsample ensemble predictions in each of the five analyses replication was identified and averaged in order to have a final unique threshold that was applied to the final continuous prediction scores. Moreover, the threshold values that generated sensitivity of 100%, 97.5%, 95%, 90%, 85%, 80%, 75% of the cross-validated train subsample ensemble predictions in each of the five analyses replication was identified, averaged, and applied to the final continuous prediction scores.
Specificity (i.e., recall), sensitivity, positive predictive value (i.e., precision), negative predictive value, balanced accuracy and F1 score (i.e., the harmonic average of the sensitivity and positive predictive value) were calculated considering the pooled categorical predictions generated with the abovementioned thresholds, which represent the estimates of the generalized site-independent performance of the algorithm when applied to perform categorical predictions of cAD/NC in new cases, such that either the balanced-accuracy is aimed to be maximized or defined levels of sensitivity are aimed to be obtained.

Feature importance

To provide a general ranking of the importance of the predictors used in this study, we applied the same five train/test split protocol to iteratively develop logistic regression models using only a single feature, in the train subsets, and these models were applied to generate the continuous prediction scores in the five test subsamples. The scores of the test subsamples were finally pooled together and used to calculate the whole sample test AUROC for each predictor. This gives a metric of importance for each predictor that is independent from both the machine learning technique used and all other predictors inserted in the algorithm. The 95% confidence interval (CI) of also these AUROCs was calculated with a stratified bootstrap procedure, with 10000 resamples and applying the bias-corrected and accelerated (BCa) approach (Efron 1987).

Results

Descriptive statistics of each feature in the cAD and NC groups are reported in Table 2. Statistics of continuous features are reported before the standardization was applied.

Feature transformation and selection

Two groups of features correlated above the 0.75 threshold were identified, respectively the three ADAS scores (ADAS11, ADAS13, ADASQ4) and two of the RAVLT scores (RAVLT-F, RAVLT-PF). Such evidence equally resulted in all the five training subsets. In all the 5 subsets, only the first principal component of each group had an eigenvalue >= 1, and these were used to substitute the correlated features as predictors (ADAS-PC1, RAVLT-F-PC1).
Across the five training subsamples used in the analyses, each feature selection procedure selected only partially overlapping subsets of relevant features, as reported in Table 4. Thus, the feature sets 2, 3, and 4 used in the analyses were in part different across the training subsamples used in the five repetitions of the analyses. This evidence further justifies our choice of creating several site-independent train and test subsamples instead of just a single training and test split, in order to provide a better and more stable estimate of the generalized performance of the algorithm.
Among the features, CDRSB, ADAS-PC1, RAVLT-I, RAVLT-F-PC1, TMTBT, and FAQ, were selected by all the three feature selection strategies in all the five repetitions of the analyses, the subtype of MCI was discarded only once, LDT twice, RAVLT-L three times and MMSE four times. All the sociodemographic characteristics were all discarded at least 6 up to 11 times out of the 15 feature sets identified in the analyses.

Performance of the predictive algorithm

The cross-validated AUROC results for each of the 52 models developed in each repetitions are reported in the supplementary data (Table S1), which ranged from a minimum value of 0.83 to a maximum value of 0.90

for the models developed with feature set 1, from 0.84 to 0.90 for the models developed with feature set 2, from 0.84 to 0.89 for the models developed with feature set 3, and from 0.83 to 0.90 for the models developed with feature set 4. These results indicate a narrow difference of performance among different feature sets, as well as among different replications and techniques, which included simple linear models such LR and NB as well as ensembling technique such as RF and GBM. The cross-validated AUROC of the weighted rank average ensembling strategy in each fold is also reported in Table S1, which ranged from a minimum of 0.86 to a maximum of 0.89.

When the test continuous prediction scores obtained with the ensembling approach were pooled, the whole sample test AUROC resulted 0.88 (95% bootstrap CI 0.85-0.91), which is plotted in Figure 1. Considering the categorical predictions generated with the threshold that maximized the training balanced accuracy, results indicated a sensitivity/recall of 77.7%, a specificity of 79.9%, a positive predictive value/precision of 68.3%, a negative predictive value of 86.5%, a balanced accuracy of 0.79, and F1-score of 0.73. Results generated applying the other thresholds are reported in Table 5.

All these results provide an estimate of the generalized performance of the algorithm when applied in new subjects which were not included in the sample used to develop the model and that have been evaluated in distinct recruiting sites.

On the server we employed in our study, training took around 12 hours for each of the 5 test folds, with a total training time of 2 days and a half. Instead, the computational time necessary to calculate the prediction using the ensemble of machine learning algorithms is less then 1 second for each case in each fold.

Importance of predictors

The AUROC of each of the various features obtained by pooling the results in the five test subsamples is reported in Table 6, ranked from the highest to the lowest AUROC, and in Figure 2, subdivided based on type of the features (i.e., sociodemographic, subtype of MCI, clinical, and neuropsychological tests). These represent an estimate of the

generalized predictive performance achievable using each feature singularly.
Sociodemographic characteristics resulted the least relevant, with age being the sole with a statistically significant AUROC (lower bound of the 95% bootstrap CI higher than 0.50) even if quite small in magnitude ($AUROC_{age}$ = 0.57). Instead, both subtypes of MCI and CDRSB demonstrated a better predictive performance ($AUROC_{MCI}$ = 0.66; $AUROC_{CDRSB}$ = 0.70), and FAQ a high AUROC of 0.78. Among the neuropsychological test scores, some of them also proved to have a high predictive capability even when used as individual predictors. The ADAS-PC1 achieved an AUROC of 0.81, RAVLT-I of 0.78, and LDT of 0.77. All other neuropsychological test scores resulted with an inferior AUROC (minimum AUROC: $AUROC_{TMTBT}$ = 0.66).
Of notice, the most relevant of the predictors, e.g., ADAS-PC1, resulted having a significantly lower test AUROC than the one demonstrated by the algorithm we developed (higher bound of the 95% bootstrap CI of ADAS-PC1 = 0.84 < lower bound of the 95% bootstrap CI of the algorithm = 0.85).

Discussion

The aim of the current study was to develop a new machine learning algorithm to allow a three-year prediction for conversion to AD in subjects diagnosed with MCI.
Considering an imminent necessity of being able to discriminate which MCI subjects will progress to AD from those who will not, as soon as in a few years the first effective treatments will be probably available (Liu, Hlávka et al. 2017), our algorithm has been designed to be used as a prognosis support tool for MCI patients, which is cost-effective and easily translatable to clinical practice. This would allow timely planning of early interventions for such individuals. Further, our algorithm can be employed as a tool during the recruitment of MCI subjects for clinical trials which aim to investigate innovative treatments of AD. The opportunity to recruit only subjects at true risk of future conversion to AD - who most likely show the earliest brain changes underlying AD

pathology – will drastically reduce the costs to run such clinical trials and result in improved outcomes.
In contrast with many of the machine learning approaches that have been previously presented, our algorithm aimed to achieve good predictive performance based only on a reduced set of sociodemographic characteristics, clinical information, and neuropsychological tests scores. It does not rely on information coming from procedures that are currently still expensive, invasive, or not widespread available in many clinical settings, such as neuroimaging techniques, lumbar puncture, and genetic testing.
The algorithm was developed using a sample of MCI subjects recruited in the ADNI study and we applied a site-independent testing protocol in order to obtain results which represent a better estimate of the expected performance when the algorithm is applied in distinct clinical centers. To the best of our knowledge, this is the first algorithm that was tested ensuring independence between the train and test sets regarding the sites where the subjects were recruited from.
Even using such a rigid testing protocol, the algorithm demonstrated a high predictive performance, showing a test AUROC of 0.88, a sensitivity of 77.7%, and a specificity of 79.9% when the classification threshold was optimized to achieve the best possible balanced accuracy. Of particular interest is the achievement of 40.2%/53% specificity and 48.3%/53% positive predictive value when the threshold was further optimized to achieve a sensitivity of respectively 100% and 95%. These results support the utility of our algorithm especially as a potential screening tool, i.e., an algorithm that can provide a marginal number of false negative predictions at the cost of a higher number of false positives. Thus, our algorithm would turn out to be particularly useful in case another more accurate, and especially more sensitive tool will become available, however which requires additional expensive or invasive-to-collect information. In such case, our algorithm can be used as a first step to significantly reduce the number of subjects which require examination using more precise, yet less easily applicable procedures at a later stage. Considering an expected conversion rate of 20%-40% from MCI to AD in three years, the expected percentage of subjects confidently predicted as non-converters would be estimated as

being 32%-24% subsequently, leaving only the remaining 68%-76% of subjects with the necessity of further investigations.

Making a proper comparison of our algorithm with all others previously published is not a trivial task, especially considering the different and reduced level of independent validation most of these algorithms have undergone so far.

In some studies, algorithms which used as predictive information some type of functional brain imaging, such as PET and fMRI, and/or CSF investigations demonstrated particularly high cross-validated performance, with AUROCs close to 0.95 (Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017). A recent study presented an algorithm based on regional information from a single amyloid PET scan which demonstrated a test performance of an AUROC of 0.91 and an unbalanced accuracy of 0.84 in the ADNI sample for a prediction of conversion in 2 years (Mathotaarachchi, Pascoal et al. 2017), thus showing a higher predictive performance than what was achieved by our algorithm.

In addition, some studies which used only structural MRI also demonstrated high cross-validated (i.e., Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017): AUROC = 0.932; balanced accuracy = 0.886) and nested cross-validated performance (sensitivity = 85%; specificity = 84.78%; Guo, Lai et al. 2017). Similarly, high cross-validated results were found by other studies who combined structural MRI with clinical and neuropsychological information (e.g., Plant, Teipel et al. 2010, Apostolova, Hwang et al. 2014, Clark, Kapur et al. 2014, Agarwal, Ghanty et al. 2015, Moradi, Pepe et al. 2015, Clark, McLaughlin et al. 2016, Dukart, Sambataro et al. 2016, Hojjati, Ebrahimzadeh et al. 2017, Long, Chen et al. 2017, Minhas, Khanum et al. 2017): AUROC = 0.902; balanced accuracy = 80.5%) In addition, a recent study (Spasov, Passamonti et al. 2019) presented a highly performing deep learning algorithm (AUROC = 0.925; accuracy = 86%; sensitivity = 87.5%; specificity = 85%) and, to the best of our knowledge, this is the only available study using structural MRI in which a proper testing of the algorithm was performed.

Some particularly promising cross-validated results were also found in some studies which considered also APOE genotyping, together with EEG, (Vecchio, Miraglia et al. 2018): AUROC = 0.97; sensitivity = 96.7%;

specificity = 86%) or blood biomarkers (Apostolova, Hwang et al. 2014, Agarwal, Ghanty et al. 2015, Dukart, Sambataro et al. 2016): balanced accuracy = 92.5%). Thus, the use of brain imaging, CSF, and/or other biomarkers as predictive information may have, to some degree, resulted in a better predictive performance compared to our algorithm, which did not use any of these types of information.

While the results of the previous studies indicate that neuroimaging biomarkers hold great promise for predicting conversion to AD, the performance increase gained by including biomarker information is questioned and much debated (Fleisher, Sun et al. 2008, Johnson, Vandewater et al. 2014, Clark, McLaughlin et al. 2016). Instead, neuropsychological measures of cognitive functioning are possibly equally excellent predictors of progression to dementia. For example, in a study by Fleisher and colleagues, common cognitive tests provide better predictive accuracy than imaging measures for predicting progression to AD in subject with moderate stages of amnestic MCI (Fleisher, Sun et al. 2008), and in another study by Clark and colleagues, models developed using only socio-demographic information, clinical information and neuropsychological test scores (focusing on verbal fluency scores) resulted in an AUROC score of 0.87 and a balanced accuracy of 0.84, while including brain imaging did not significantly improve this performance (AUROC = 0.81, accuracy = 0.83) (Clark, McLaughlin et al. 2016).

Moreover, the cost of the standard procedure in the clinical process of diagnosing AD (which entails the clinical consultation, including the patient's administrative admission, anamnesis, physical examination, neuropsychological testing, test evaluation and diagnosis conference & physician letter) is relatively low at an estimated 110 € (US$115) on average, while the use of additional advanced technical procedures, such as blood sampling, CT, MRI, PET & CSF procedures, which are required following deficits in neuropsychological test results and depends on the patient's suspected diagnosis of MCI, AD or other dementia types (which is increasingly associated with higher frequencies of using cost-intensive imaging & CSF procedures), drives costs up to 649 € (US$676) in case of an AD diagnosis according to a study in a German memory clinic (Michalowsky, Flessa et al. 2017).

In this regard, the use of advanced technological procedures, rather than clinical consultation and neuropsychological testing, is driving costs in the diagnostic process and as such, will also increase the costs of predictive algorithms based on information of imaging, blood sampling or CSF procedures compared to those algorithms that rely only on sociodemographic, clinical, and neuropsychological predictive information, like the one we present in this study. In addition, even if nowadays some forms of neuroimaging investigations are often routinely performed, for example in order to evaluate other potential comorbidities such as neurovascular problems or regional atrophies, and thus such information may result already available without additional costs, a clear evidence of its relevance to improve predictions based only on neuropsychological and clinical measures is still lacking, as it has already been discussed above, and still requires further investigations.

Additionally, our algorithm demonstrated similar predictive performance compared to other top-performing algorithms based only on sociodemographic, clinical, and neuropsychological predictive information. For example, in a first study by Clark and colleagues, they used only a simple cross-validation protocol to investigate the performance of their algorithm to make prediction of conversion at 1 year or more (AUROC = 0.88, balanced accuracy = 0.84) (Clark, Kapur et al. 2014), while in another study they used a more sound nested cross-validation protocol to investigate the predictive performance of their algorithm at 4 years (AUROC = 0.87, balanced accuracy = 0.79) (Clark, McLaughlin et al. 2016).

Our results originate from a proper testing protocol and represent a better unbiased estimate of the generalized performance of the algorithm. Only a very small number of machine learning algorithms for the prediction of conversion from MCI to AD were subjected to a proper testing protocol, rather than only a cross-validation protocol, which limits the soundness of the evidence of their predictive performance. As such, apart from (Mathotaarachchi, Pascoal et al. 2017, Spasov, Passamonti et al. 2019), all the previously mentioned results may be optimistically biased estimates of the generalized performance of such algorithms as a proper testing protocol was not applied.

We previously presented another machine learning algorithm that performs a prediction of conversion to AD in MCI subjects (Grassi,

Loewenstein et al. 2018, Grassi, Perna et al. 2018). However, the algorithm described here has distinct characteristics and can be considered at a more advanced stage of validation. First, the current algorithm does not require any neuroimaging information, while our previous method relied on a clinicians' rating of the atrophy in three brain structures, evaluated by observing standardized images coming from a structural magnetic resonance. Structural magnetic resonance is widespread also in clinical settings nowadays, it is less expensive than other neuroimaging evaluation such as functional magnetic resonance and positron emission tomography, and the use of a clinician-administered visual scale allows to bypass the obstacles related to the non-automatic calibration of data coming from different magnetic resonance scanners. Nevertheless, the fact that our new algorithm does not necessitate any magnetic resonance evaluation makes its use even more easily translatable in practice, and less expensive. Moreover, even though our former algorithm showed higher cross-validated performance (AUROC = 0.91, sensitivity = 86.7% and specificity = 87.4% at the best balanced accuracy; Grassi, Perna et al. 2018), a solid testing of its performance is still lacking and, at the moment, only a preliminary evidence via a transfer learning approach is available (Grassi, Perna et al. 2018). Instead, the protocol applied in the current study provides a better and sounder evaluation of the actual predictive performance of this new algorithm.

Beyond testing the algorithm's predictive accuracy, we also aimed to provide a first indication of the importance of the variables used as predictors. The opportunity to provide an explanation of how the model works and performs its prediction is crucial to foster its application in clinical practice (Perna, Grassi et al. 2018). However, given the architectural complexity of the algorithm we developed, this is not a straightforward task. Several different approaches have been proposed, all of them providing a different, and only partial explanation of an algorithm's functioning (Du, Liu et al. 2018). Thus, we decided to leave complex and more extensive investigations to a future study which will be fully dedicated to this goal. Instead, we simply investigated the predictive role of each predictor individually, which can evidence the amount of predictive information carried by each predictor. However, it does not allow to identify potential interactions among multiple

predictors that could have been modeled by the algorithm and that can relevantly contribute to its high predictive performance.
In line with the evidence in our previous study (Grassi, Perna et al. 2018), sociodemographic characteristics seem not to be particularly relevant in discriminating cAD and NC MCI subjects. Furthermore, in both studies, age was the sole of these characteristics showing a significant, even if very limited, predictive power. Also, sociodemographic characteristics resulted to be the most often discarded features by the feature selection strategies we applied in our study, once again suggesting their poor predictive relevance.
Instead, the clinical scale scores, the subtype of MCI, and the neuropsychological test scores resulted markedly predictive. Their test AUROC ranged from 0.658 to 0.809, and even the least predictive of them had a 95% CI higher than 0.6. The evidence of their predictive importance was expected. These features measure core elements of the progressive decline leading to a full manifestation of AD, such as the memory and other cognitive functions deterioration, and the consequent functional impairment.
In our algorithm, as well as in several previously presented algorithms which included clinical, and neuropsychological predictors, some of these were also reassessed at later follow-ups in order to investigate when a conversion to AD occurred after the baseline assessment. As a matter of facts, MMSE and CDR scores below certain cut-offs and a cognitive impairment in at least two cognitive domains are necessary criteria to receive a diagnosis of probable AD, evidencing a conversion from MCI to AD. Using some measures at baseline to predict the same or related measures at a future follow-up time is a strategy at the foundation of time-series analyses (i.e., autoregressive models). The same measure may result correlated to itself at different future times (i.e., autocorrelation), thus making relevant predictive information at the disposal of the predictive model. Instead, in other occasions, a measure may result uncorrelated to itself across different times of assessment. The result of a significant individual predictive performance of all neuropsychological tests, MMSE, and CDR baseline scores evidences the former in our data, and it may generally be interpreted as that the more severe is the level of impairment reached by a subject, the higher becomes the probability of its progression until a conversion to AD within

the following three years. The use of such autocorrelated information as predictors may have relevantly contributed to the high performance achieved by our as well other algorithms which included them, compared to others which did not (Chapman, McCrary et al. 2011, Battista, Salvatore et al. 2017).
Moreover, the first principal component of the three ADAS scores, which resulted in the most individually important predictor, demonstrated a test AUROC significantly lower than the one achieved by the entire algorithm. The results of our, as well as other previous studies, had already showed that machine learning algorithms can effectively be used to combine these individual pieces of information, providing a better identification of cAD among MCI subjects than what it would be possible using each of them singularly (Clark, Kapur et al. 2014, Johnson, Vandewater et al. 2014, Clark, McLaughlin et al. 2016, Grassi, Loewenstein et al. 2018).
Our study has some limitations that should be taken into account and that will be addressed in the future stages of our research. First, even if we iteratively ensured that the subjects used for testing were always recruited in different sites than those used in the development of the algorithm, it is important to note that all the ADNI recruiting sites were located in the USA or Canada. Even if this can be considered an important step forward towards the demonstration of the generalized performance of the proposed algorithm, still these sites may not be completely representative of the entire population of centers in which the algorithm may aspire to be used. Our aim was to develop an algorithm that may be applied also beyond US and Canada centers only, and perhaps also clinical centers without any research inclinations. MCI subjects referring to these extended range of centers might have peculiar characteristics and the algorithm might show reduced predictive accuracy when applied to them. In order to at least partially address this potential bias, we plan to first test and then re-optimize our algorithm using further datasets coming from the several international replications of the North American Alzheimer's Disease Neuroimaging Initiative (https://www.alz.org/research/for_researchers/partnerships/wwadni). In addition, inclusion and exclusion criteria may have excluded from ADNI, and in turn from our analyses, some MCI subjects with peculiar characteristics, e.g., MCI subjects with high level of depression or currently taking some of the medications that were excluded from the

study. Once again, the algorithm might show reduced predictive accuracy when applied to them and further testing in a less selected sample should be performed before a safe use of the algorithm can be guaranteed with these peculiar MCI subjects.
Furthermore, our final algorithm is based on an ensemble of several lower-level machine learning algorithms, including some that use the entire initial set of predictors as feature set. Thus, all predictors currently remain necessary to be assessed, even if some of them may contribute poorly or even not at all to the prediction. Although the ensembling approach we used may have effectively prevented that such irrelevant predictors decreased the algorithm accuracy, a further reduction of the amount of information necessary to be assessed and used by the algorithm would permit to reduce the costs associated with its application. At the same time, our algorithm may have missed to take into account relevant pieces of information that can improve the accuracy of its predictions.
It should be also noted that compensatory neurophysiological mechanisms, including for instance cognitive reserve factors such as bilingualism that are latent in MCI subjects, might result in misclassifications of MCI converters and non-converters (Alladi, Bak et al. 2013, Lojo-Seoane, Facal et al. 2018). It would be important to take this into account for predictive models, like ours, that exclusively relies on quantitative psychological test scores to predict the conversion to AD in MCI patients, as these compensatory brain mechanisms might not be reflected during neuropsychological testing and perhaps potentially impact the performance of the algorithm.
Finally, our algorithm currently operates three-year predictions in subjects that already manifest MCI. As the new arriving treatments are expected to be the more effective the earlier they will be started, algorithms that can perform accurate predictions at even earlier stages of deterioration than MCI, and in a longer time frame, will be of particular relevance. A preliminary attempt has already been done in our previous study (Grassi, Perna et al. 2018), employing also a sample of subjects with Pre-mild Cognitive Impairment (Chao, Mueller et al. 2010), as well as in other previous studies which developed algorithm that aimed to make predictions for periods longer than three years (Agarwal, Ghanty et al. 2015, Clark, McLaughlin et al. 2016). Future steps in our research will

take into account this necessity, exploring the opportunity of making predictions at longer time periods and in earlier-stage subjects.

Conclusions

We developed an algorithm to predict three-year conversion to AD in MCI subjects, based on a weighted rank average ensemble of several supervised machine learning algorithms. It demonstrated high predictive accuracy when tested via a sound train/test split protocol, exhibiting especially good predictive performance when the algorithm was optimized as a screening tool. Predictions are performed using only a restricted set of sociodemographic characteristics, clinical information, and neuropsychological test scores, which makes its application of easy translation into clinical practice, as well as useful in improving the recruitment of MCI subjects at true risk of conversion to AD in clinical trials.
It is important to conclude highlighting that any prediction, including those provided by machine learning algorithms, is probabilistic in its nature and always comes with a certain degree of imprecision. The advantage in the potential use of algorithmic decision-making tools is that such imprecision is defined by a known and objectively investigated degree of confidence. However, in order to guarantee such confidence, several and continuous tests of an algorithm have to be performed before its application can be safely recommended. Further tests and optimizations will follow this study in the attempt to provide additional evidence of its accuracy in generalized applications, and to improve its cost-effectiveness.

Acknowledgments: The authors would like to express their thanks to Lianne Ippel, Seun Onaopepo Adekunle, Alexander Malic, and Pedro Hernandez from the Institute of Data Science at Maastricht University for their valuable assistance, contributing their expertise to the experimental design of the analyses of this study. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

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

Availability of data and materials: Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu), which is easily available for download from the Laboratory of Neuroimaging (LONI) website to the research public.

Author contributions: MG contributed to the design of the work, the design and execution of the analyses, interpretation of results, drafting of the paper and knowledge communication. NR contributed to the design of the work and the analyses, interpretation of results and drafting of the paper, as well as general project management, knowledge utilization and communication, interactions with research professionals. DL contributed to the initial conception of the work and revising the manuscript. DC, KS, and GP contributed supervising the work and revising the manuscript. MD contributed to the design of the work and

the analyses, interpretation of results, supervising the work and revising the manuscript. All authors read and approved the final manuscript.

Conflict of interest: The authors declare that they have no competing interests.

Contribution to the Field: Alzheimer's Disease (AD) is the most common form of dementia, whose progression comprises a long, unnoticed preclinical stage, followed by a prodromal stage of Mild Cognitive Impairment (MCI) that leads to severe dementia and eventually death. As currently available and emerging therapies likely have the greatest impact when provided at the earliest disease stage, the prompt identification of subjects at high risk for conversion to AD is of great importance both to improve the recruitment of MCI subjects in clinical trials and considering that the available and emerging therapies likely have the greatest impact when provided at the earliest disease stage. Considerable effort has been dedicated in developing predictive algorithms, but they generally rely on expensive and/or invasive predictors, such as brain imaging or CSF biomarkers, usually serve the purpose of a proof-of-concept. In this work, we present a new machine learning predictive algorithm to identify MCI subjects at risk for conversion to AD in the following three years. The algorithm demonstrated a high prognostic performance even if it is based only on non-invasively and easily in-the-clinic collectable predictors (i.e., sociodemographic and clinical characteristics, neuropsychological test scores), which facilitates its potential translation into practice.

References

Agarwal, S., P. Ghanty and N. R. Pal (2015). "Identification of a small set of plasma signalling proteins using neural network for prediction of Alzheimer's disease." Bioinformatics **31**(15): 2505-2513.

Alladi, S., T. H. Bak, V. Duggirala, B. Surampudi, M. Shailaja, A. K. Shukla, J. R. Chaudhuri and S. Kaul (2013). "Bilingualism delays age at onset of dementia, independent of education and immigration status." Neurology **81**(22): 1938-1944.

Altman, N. S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." Am. Stat. **46**(3): 175-185.

Alzheimer's, A. (2018). "2018 Alzheimer's disease facts and figures." Alzheimers. Dement. **14**(3): 367-429.

Alzheimer's Disease, I. (2018). World Alzheimer Report 2018 The state of the art of dementia research: New frontiers.

Apostolova, L. G., K. S. Hwang, O. Kohannim, D. Avila, D. Elashoff, C. R. Jack, Jr., L. Shaw, J. Q. Trojanowski, M. W. Weiner, P. M. Thompson and I. Alzheimer's Disease Neuroimaging (2014). "ApoE4 effects on automated diagnostic classifiers for mild cognitive impairment and Alzheimer's disease." Neuroimage Clin **4**: 461-472.

Battista, P., C. Salvatore and I. Castiglioni (2017). "Optimizing Neuropsychological Assessments for Cognitive, Behavioral, and Functional Impairment Classification: A Machine Learning Study." Behav. Neurol. **2017**: 1850909.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Chao, L. L., S. G. Mueller, S. T. Buckley, K. Peek, S. Raptentsetseng, J. Elman, K. Yaffe, B. L. Miller, J. H. Kramer, C. Madison, D. Mungas, N. Schuff and M. W. Weiner (2010). "Evidence of neurodegeneration in brains of older adults who do not yet fulfill MCI criteria." Neurobiol Aging **31**(3): 368-377.

Chapman, R. M., J. W. McCrary, M. N. Gardner, T. C. Sandoval, M. D. Guillily, L. A. Reilly and E. DeGrush (2011). "Brain ERP components predict which individuals progress to Alzheimer's disease and which do not." Neurobiology of Aging **32**(10): 1742-1755.

Clark, D. G., P. Kapur, D. S. Geldmacher, J. C. Brockington, L. Harrell, T. P. DeRamus, P. D. Blanton, K. Lokken, A. P. Nicholas and D. C. Marson (2014). "Latent information in fluency lists predicts functional decline in persons at risk for Alzheimer disease." Cortex **55**: 202-218.

Clark, D. G., P. M. McLaughlin, E. Woo, K. Hwang, S. Hurtz, L. Ramirez, J. Eastman, R.-M. Dukes, P. Kapur, T. P. DeRamus and L. G. Apostolova (2016). "Novel verbal fluency scores and structural brain imaging for prediction of cognitive outcome in mild cognitive impairment." Alzheimers. Dement. **2**: 113-122.

Devanand, D. P., X. Liu, M. H. Tabert, G. Pradhaban, K. Cuasay, K. Bell, M. J. de Leon, R. L. Doty, Y. Stern and G. H. Pelton (2008). "Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease." Biol. Psychiatry **64**(10): 871-879.

Du, M., N. Liu and X. Hu (2018). "Techniques for Interpretable Machine Learning." arXiv [cs.LG].

Dukart, J., F. Sambataro and A. Bertolino (2016). "Accurate Prediction of Conversion to Alzheimer's Disease using Imaging, Genetic, and Neuropsychological Biomarkers." J. Alzheimers. Dis. **49**(4): 1143-1159.

Efron, B. (1987). "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association **82**(397): 171-185.

Fleisher, A. S., S. Sun, C. Taylor, C. P. Ward, A. C. Gamst, R. C. Petersen, C. R. Jack, Jr., P. S. Aisen and L. J. Thal (2008). "Volumetric MRI vs clinical predictors of Alzheimer disease in mild cognitive impairment." Neurology **70**(3): 191-199.

Folstein, M. F., S. E. Folstein and P. R. McHugh (1975). "“Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician." J. Psychiatr. Res.

Giannakopoulos, P. (2017). "Alzheimer disease biomarkers: Facing the complexity." J. Alzheimers Dis. Parkinsonism **07**(04).

Grassi, M., D. A. Loewenstein, D. Caldirola, K. Schruers, R. Duara and G. Perna (2018). "A clinically-translatable machine learning algorithm for the prediction of Alzheimer’s disease conversion: further evidence of its accuracy via a transfer learning approach." Int. Psychogeriatr. **14**: 1-9.

Grassi, M., G. Perna, D. Caldirola, K. Schruers, R. Duara and D. A. Loewenstein (2018). "A Clinically-Translatable Machine Learning Algorithm for the Prediction of Alzheimer’s Disease Conversion in Individuals with Mild and Premild Cognitive Impairment." Journal of Alzheimer's Disease **61**(4): 1555-1573.

Guo, S., C. Lai, C. Wu, G. Cen and I. The Alzheimer's Disease Neuroimaging (2017). "Conversion Discriminative Analysis on Mild Cognitive Impairment Using Multiple Cortical Features from MR Images." Front. Aging Neurosci. **9**.

Hojjati, S. H., A. Ebrahimzadeh, A. Khazaee, A. Babajani-Feremi and I. Alzheimer's Disease Neuroimaging (2017). "Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM." J. Neurosci. Methods **282**: 69-80.

Johnson, P., L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, L. S. Macaulay, K. A. Ellis, C. Szoeke, R. N. Martins, C. C. Rowe, C. L. Masters, D. Ames and P. Zhang (2014). "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease." BMC Bioinformatics **15**(16): S11.

Kingma, D. P. and J. Ba (2014). "Adam: A Method for Stochastic Optimization." arXiv [cs.LG].

Liu, J. L., J. P. Hlávka, R. Hillestad and S. Mattke (2017). "Assessing the preparedness of the US health care system infrastructure for an Alzheimer's treatment." Available at:) The RAND Corporation, Santa Monica, CA.

Lojo-Seoane, C., D. Facal, J. Guàrdia-Olmos, A. X. Pereiro and O. Juncos-Rabadán (2018). "Effects of Cognitive Reserve on Cognitive Performance in a Follow-Up Study in Older Adults With Subjective Cognitive Complaints. The Role of Working Memory." Frontiers in Aging Neuroscience **10**.

Long, X., L. Chen, C. Jiang, L. Zhang and I. Alzheimer's Disease Neuroimaging (2017). "Prediction and classification of Alzheimer disease based on quantification of MRI deformation." PLoS One **12**(3): e0173372.

Mason, L., J. Baxter, P. L. Bartlett and M. R. Frean (2000). Boosting algorithms as gradient descent. Advances in neural information processing systems.

Mathotaarachchi, S., T. A. Pascoal, M. Shin, A. L. Benedet, M. S. Kang, T. Beaudry, V. S. Fonov, S. Gauthier, P. Rosa-Neto and I. Alzheimer's Disease Neuroimaging (2017). "Identifying incipient dementia individuals using machine learning and amyloid imaging." Neurobiol Aging **59**: 80-90.

Michalowsky, B., S. Flessa, J. Hertel, O. Goetz, W. Hoffmann, S. Teipel and I. Kilimann (2017). "Cost of diagnosing dementia in a German memory clinic." Alzheimers. Res. Ther. **9**(1): 65.

Minhas, S., A. Khanum, F. Riaz, A. Alvi and S. A. Khan (2017). "A Nonparametric Approach for Mild Cognitive Impairment to AD Conversion Prediction: Results on Longitudinal Data." IEEE J Biomed Health Inform **21**(5): 1403-1410.

Moradi, E., A. Pepe, C. Gaser, H. Huttunen, J. Tohka and I. Alzheimer's Disease Neuroimaging (2015). "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects." Neuroimage **104**: 398-412.

Morris, J. C. (1993). "The Clinical Dementia Rating (CDR): current version and scoring rules." Neurology **43**(11): 2412-2414.

Opitz, D. and R. Maclin (1999). "Popular Ensemble Methods: An Empirical Study." 1 **11**: 169-198.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Perna, G., M. Grassi, D. Caldirola and C. Nemeroff (2018). "The revolution of personalized psychiatry: will technology make it happen sooner?" Psychological medicine **48**(5): 705-713.

Petersen, R. C., J. E. Parisi, D. W. Dickson, K. A. Johnson, D. S. Knopman, B. F. Boeve, G. A. Jicha, R. J. Ivnik, G. E. Smith, E. G. Tangalos, H. Braak and E. Kokmen (2006). "Neuropathologic features of amnestic mild cognitive impairment." Arch Neurol **63**(5): 665-672.

Pfeffer, R. I., T. T. Kurosaki, C. H. Harrah, Jr., J. M. Chance and S. Filos (1982). "Measurement of functional activities in older adults in the community." J. Gerontol. **37**(3): 323-329.

Plant, C., S. J. Teipel, A. Oswald, C. Bohm, T. Meindl, J. Mourao-Miranda, A. W. Bokde, H. Hampel and M. Ewers (2010). "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease." Neuroimage **50**(1): 162-174.

Platt, J. C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." ADVANCES IN LARGE MARGIN CLASSIFIERS: 61-74.

Python Software, F. Python Language.

Reitan, R. M. (1958). "Validity of the Trail Making Test as an Indicator of Organic Brain Damage." Percept. Mot. Skills **8**(3): 271-276.

Rish, I. and Others (2001). An empirical study of the naive Bayes classifier. IJCAI 2001 workshop on empirical methods in artificial intelligence. **3:** 41-46.

Roberts, R. O., D. S. Knopman, M. M. Mielke, R. H. Cha, V. S. Pankratz, T. J. H. Christianson, Y. E. Geda, B. F. Boeve, R. J. Ivnik, E. G. Tangalos, W. A. Rocca and R. C. Petersen (2014). "Higher risk of progression to dementia in mild cognitive impairment cases who revert to normal." Neurology **82**(4): 317-325.

Rosen, W. G., R. C. Mohs and K. L. Davis (1984). "A new rating scale for Alzheimer's disease." Am. J. Psychiatry **141**(11): 1356-1364.

Schmidt, M. and Others (1996). Rey auditory verbal learning test: A handbook, Western Psychological Services Los Angeles, CA.

Schölkopf, B., A. J. Smola and F. Bach (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press.

Spasov, S., L. Passamonti, A. Duggento, P. Liò and N. Toschi (2019). "A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease." Neuroimage **189**: 276-287.

Sperling, R. and K. Johnson (2013). "Biomarkers of Alzheimer disease: current and future applications to diagnostic criteria." Continuum (Minneap Minn) **19**(2 Dementia): 325-338.

Vecchio, F., F. Miraglia, F. Iberite, G. Lacidogna, V. Guglielmi, C. Marra, P. Pasqualetti, F. D. Tiziano and P. M. Rossini (2018). "Sustainable method for Alzheimer dementia prediction in mild cognitive impairment: Electroencephalographic connectivity and graph theory combined with apolipoprotein E." Ann. Neurol. **84**(2): 302-314.

Wechsler, D. (1945). "Wechsler Memory Scale." PsycTESTS Dataset.

Wechsler, D. (1997). WMS-III: Wechsler memory scale administration and scoring manual, Psychological Corporation.

Wolpert, D. H. (1992). "Stacked generalization." Neural Netw. **5**(2): 241-259.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the royal statistical society: series B (statistical methodology) **67**(2): 301-320.

Table 1. Abbreviations of neuropsychological tests.

ADAS11	Cognitive Subscale (11 items) Alzheimer's Disease Assessment Scale
ADAS13	Cognitive Subscale (13 items) Alzheimer's Disease Assessment Scale
ADASQ4	Task 4 of the Cognitive Subscale (11 items) Alzheimer's Disease Assessment Scale
CDRSB	Sum of Boxes score of the Clinical Dementia Rating Scale;
DIGIT	Digit Span Test score
FAQ	Functional Activities Questionnaire
LDT	Logic Memory subtest of the of the Wechsler Memory Scale-Revised
RAVLT	Rey Auditory Verbal Learning Test
RAVLT-F	Forgetting score of the Rey Auditory Verbal Learning Test
RAVLT-I	Immediate score of the Rey Auditory Verbal Learning Test
RAVLT-L	Learning score of the Rey Auditory Verbal Learning Test
RAVLT-PF	Percent forgetting score of the Rey Auditory Verbal Learning Test
TMTBT	Trial Making Test, version B

Table 2. Descriptive statistics

Continuous predictors	Non-converters		Converters		Missing values	
	Mean	S.D.	Mean	S.D.	N	%
Age	72.42	7.54	74.19	6.88	/	/
Years of education	16.18	2.74	15.74	2.83	/	/
CDRSB	1.26	0.70	1.95	1.01	/	/
ADAS11	8.67	3.78	12.94	4.26	1	0.18%
ADAS13	13.89	5.81	21.05	5.72	3	0.55%
ADASQ4	4.61	2.35	7.16	2.04	/	/
MMSE	28.01	1.71	26.85	1.72	/	/
RAVLT-I	37.84	10.47	28.05	6.74	/	/
RAVLT-L	4.76	2.59	2.90	2.11	/	/
RAVLT-F	4.37	2.46	5.20	2.30	/	/
RAVLT-PF	51.09	30.92	78.20	28.04	/	/
LDT	6.84	3.12	3.59	2.89	/	/
DIGIT	40.24	10.42	34.86	11.02	290	52.73%
TMTBT	100.30	49.56	141.24	79.66	4	0.73%
FAQ	1.76	2.75	5.81	5.00	4	0.73%

Categorical predictors		Non-converters		Converters		Missing values	
		N	%	N	%	N	%
Sex	Male	220	62.32%	118	59.90%	/	/
	Female	133	37.68%	79	40.10%		
Subtype of MCI	Early	196	47.88%	22	11.17%	/	/
	Late	184	52.12%	175	88.83%		
Marital status	Never married	6	1.70%	3	1.52%	3	0.55%
	Married	267	75.64%	161	81.73%		
	Divorced	35	9.92%	13	6.60%		
	Widowed	42	11.90%	20	10.15%		

S.D = Standard Deviation; N = numbers of subjects.

Table 3. Allocation of the ADNI study recruitment sites in the five subsets.

Subset	Recruiting Sites																		Non-converters		Converters	
																			N	%	N	%
A	6	12	18	21	126	127	128	137											74	64.35%	41	35.65%
B	2	3	9	23	24	29	36	37	94	99	114								70	64.22%	39	35.78%
C	7	13	14	33	41	67	73	98	100	109	116								70	64.22%	39	35.78%
D	5	16	22	27	31	35	123	130	141	153									70	64.22%	39	35.78%
E	10	11	19	32	51	52	53	57	62	68	72	82	129	131	133	135	136	941	69	63.89%	39	36.11%

The code of the recruitment sites follows the coding convention used in the ADNI study.

Table 4. Feature sets 2, 3, and 4 in each of the five replications of the analyses.

Predictors	Feature Set 2					Feature Set 3					Feature Set 4				
	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Age	x	x		x	x						x	x	x	x	x
Years of education											x	x	x	x	
CDRSB	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ADAS-PC1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
MMSE	x	x	x	x	x	x			x		x	x	x	x	
RAVLT-I	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
RAVLT-L	x	x	x	x	x	x			x		x	x	x	x	x
RAVLT-F-PC1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
LDT	x	x	x	x	x		x		x	x	x	x	x	x	x
TMTBT	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
FAQ	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Sex						x	x	x	x	x	x	x	x	x	
Subtype of MCI	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Marital Status - Never Married						x	x		x	x	x				
Marital Status - Married						x	x	x	x	x	x	x			
Marital Status - Divorced						x	x	x	x	x	x	x			
Marital Status - Widowed						x	x	x	x	x	x	x			

A-E indicates the 5 independent subsets in which the analyses have been replicated.

Table 5. Test performance of the algorithm.

Performance Level (Target)	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Balanced Accuracy	F1-score
Sensitivity of 1	100%	40.2%	48.3%	100.0%	0.701	0.651
Sensitivity of .975	97.5%	49.6%	51.9%	97.2%	0.735	0.677
Sensitivity of .95	94.9%	53.0%	53.0%	94.9%	0.739	0.680
Sensitivity of .90	88.8%	67.4%	60.3%	91.5%	0.781	0.719
Sensitivity of .85	84.3%	73.1%	63.6%	89.3%	0.787	0.725
Sensitivity of .80	79.2%	79.6%	68.4%	87.3%	0.794	0.734
Sensitivity of .75	71.6%	84.1%	71.6%	84.1%	0.779	0.716
Best Balanced Accuracy	77.7%	79.9%	68.3%	86.5%	0.788	0.727

Table 6. Individual test pooled AUROC of each feature.

	AUROC	95% Bootstrap CI	
ADAS-PC1	0.809	0.772	0.842
RAVLT-I	0.777	0.737	0.814
FAQ	0.777	0.733	0.816
LDT	0.770	0.726	0.808
RAVLT-L	0.707	0.661	0.750
CDRSB	0.697	0.648	0.740
RAVLT-F-PC1	0.685	0.639	0.730
MMSE	0.678	0.631	0.723
Subtype of MCI	0.658	0.610	0.702
TMTBT	0.658	0.608	0.704
Age	0.564	0.511	0.614
Years of education	0.540	0.494	0.590
Marital Status - Married	0.506	0.452	0.547
Marital Status - Divorced	0.501	0.449	0.543
Marital Status - Never Married	0.488	0.439	0.537
Marital Status - Widowed	0.487	0.430	0.529
Sex	0.475	0.413	0.512

AUROC = Area Under the Receiving Operating Curve.

Figure 1. Area Under the Receiving Operating Curve of the pooled test predictions.

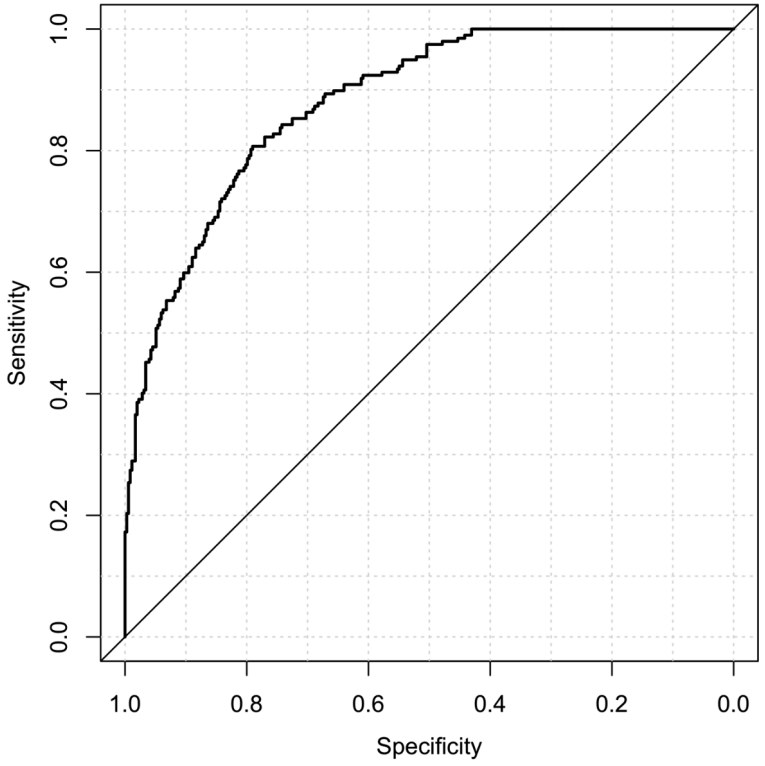
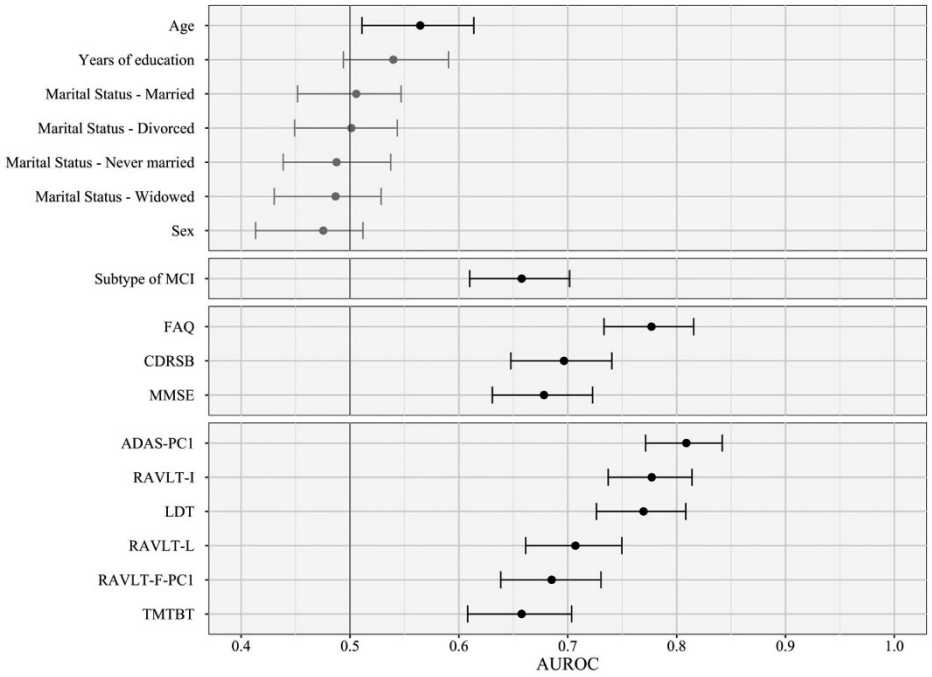


Figure 2. Area Under the Receiving Operating Curve of Individual Predictors.



The figure indicates the pooled test AUROC and its 95% bootstrap CI when prediction is made considering each predictor singularly. Predictors are grouped according to conceptual domains, which in descending order are sociodemographic characteristics, subtype of MCI, clinical scale scores, and neuropsychological test scores. Non-significant AUROC (i.e., the lower bound of the CI is lower than or equal to 0.5) are in grey, significant ones in black.

CHAPTER 5

PREDICTION OF ILLNESS REMISSION IN PATIENTS WITH OBSESSIVE-COMPULSIVE DISORDER WITH SUPERVISED MACHINE LEARNING

Massimiliano Grassi[1,2], Judith Rickelt[3,4], Daniela Caldirola[1,2], Merijn Eikelenboom[5], Patricia van Oppen[5], Michel Dumontier[6], Giampaolo Perna[1,2,3,8], Koen Schruers[3]

1 Department of Clinical Neurosciences, Hermanas Hospitalarias, Villa San Benedetto Menni Hospital, Albese con Cassano, Como, Italy.
2 Department of Biomedical Sciences, Humanitas University, Rozzano, Milan, Italy.
3 Research Institute of Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands.
4 Institute for Mental Health Care Eindhoven (GGzE), Eindhoven, the Netherlands
5 Amsterdam UMC, location VUmc, Department of Psychiatry, Amsterdam Public Health research institute and GGZ inGeest Specialized Mental Health Care, the Netherlands
6 Institute of Data Science, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands.
7 Department of Psychiatry and Behavioral Sciences, Leonard Miller School of Medicine, University of Miami, FL, USA.

Reference: Grassi M, Rickelt J, Caldirola D, Eikelenboom M, van Oppen P, Dumontier M, Perna G, Schruers K. Prediction of illness remission in patients with Obsessive-Compulsive Disorder with supervised machine learning. J Affect Disord. 2022 Jan 1;296:117-125.

Abstract

Introduction. The course of OCD differs widely between individual OCD patients, varying from severe chronic symptoms to full remission. No tools for individual prediction of OCD remission are currently available. The present study aimed to develop and test a machine learning algorithm to predict OCD remission after 2 years, using solely predictors easily accessible in the daily clinical routine.
Methods. Subjects were recruited in a longitudinal multi-center study (NOCDA). Gradient boosted decision trees (GBDT) were used as a supervised machine learning technique. The training of the algorithm was performed with 227 features and 213 cases recruited in a single clinical center. Hyper-parameter optimization was performed with 10-fold cross-validation and a Bayesian optimization strategy. The predictive performance of the algorithm was subsequently tested using an independent sample of 215 cases recruited from five different centers. Between-center differences were investigated with a bootstrap resampling approach.
Results: The average predictive performance of the algorithm in the five test centers resulted in an AUROC of 0.7820, a sensitivity of 73.42%, and a specificity of 71.45%. However, a large between-center variation was observed, which was partially statistically significant even after a conservative Bonferroni correction for multiple comparisons.
Discussion. The present study developed an algorithm for OCD course prediction and subsequently tested it in different independent test samples. Although the algorithm resulted in a moderate average predictive performance, results showed a large variation in the predictive performance when tested per center. Future studies will focus on increasing the stability of the predictive performance across clinical settings, as well as on improving the overall accuracy of the algorithm.

Keywords: Obsessive-Compulsive Disorder, Machine Learning, Prognosis, Remission, Personalized Medicine.

Introduction

Obsessive-compulsive disorder (OCD) is a debilitating disorder characterized by intrusive thoughts or images (obsessions) and ritualized stereotypic and often repetitive behavior (compulsions) that are time-consuming and interfere with daily functioning (American Psychiatric Association 2013). It is listed as the tenth most disabling medical disorder in the World Health Organization (WHO) burden of disease study (Ezzati, Lopez et al. 2004) and is associated with diminished quality of life (Coluccia, Fagiolini et al. 2016, Pozza, Lochner et al. 2018)

Despite the effectiveness of selective serotonin reuptake inhibitors and cognitive-behavioral therapy, obsessive-compulsive symptoms persist in a large group of patients. Remission rates vary from 50 to 80% depending on treatment modality and definition of treatment outcome (Fineberg, Brown et al. 2012, Ost, Havnen et al. 2015, Agne, Tisott et al. 2020). OCD tends to run a chronic course in the majority of patients. Long-term treatment follow-up studies found varying remission rates of 50% to 65% (van Oppen, van Balkom et al. 2005, Kempe, van Oppen et al. 2007, Cherian, Math et al. 2014, Nakajima, Matsuura et al. 2018) with relapse during follow-up in more than half of the remitted OCD patients (Kempe, van Oppen et al. 2007). Results of long-term naturalistic studies vary widely due to differences in outcome definition and methodology. In summary, 10-30% of the OCD patients achieve complete recovery and about 25% suffer from chronic persisting or deteriorating symptoms, while the majority of the OCD patients experience partial improvement over the years, and more than half of the remitted patients subsequently relapse (Skoog, 1999 #74;Eisen, 2013 #66;Garnaat, 2015 #67}.

In sum, the course of OCD varies widely among different individuals. Several studies investigated factors associated with treatment outcome and course of OCD with the aim of finding predictors for remission, relapse, and chronicity of obsessive-compulsive symptoms. Several hypotheses including various factors such as OCD symptom severity, OCD symptom dimensions, course, insight, comorbidities, OCD-related cognitions, or social circumstances were investigated. However, results are inconclusive, often contradictory, and mostly not replicated (Keeley, Storch et al. 2008, Knopp, Knowles et al. 2013, Hazari, Narayanaswamy et al. 2016). Thus, the possibility of making a prompt individual-level

prediction of the clinical course of OCD is currently limited because reliable clinically relevant predictors are not available (Schuurmans, van Balkom et al. 2012, Knopp, Knowles et al. 2013, Hazari, Narayanaswamy et al. 2016)

In addition, different factors may contribute to the prognosis of OCD and thus predictions based on single factors are too restricted and inaccurate to be used in clinical practice. Instead, models that simultaneously exploit the information coming from several potential predictors may achieve a better predictive capability.

Machine learning (ML) techniques can be used to create precisely such models. ML techniques use known training examples to create algorithms able to provide the best possible prediction when applied to new cases whose outcome is still unknown. It is a fast-growing field at the crossroads of computer science, engineering, and statistics "that gives computers the ability to learn without being explicitly programmed" (Samuel 1959).

A few attempts to apply such techniques to achieve clinically relevant predictions in OCD patients have already been made (Salomoni, Grassi et al. 2009, Hoexter, Miguel et al. 2013, Askland, Garnaat et al. 2015, Yun, Jang et al. 2015, Mas, Gasso et al. 2016, Lenhard, Sauer et al. 2018, Reggente, Moody et al. 2018, Agne, Tisott et al. 2020, Metin, Balli Altuglu et al. 2020). Although some of the algorithms showed high preliminary predictive accuracy, they have remained just proofs-of-concept, with a lack of any testing in further independent samples. Evidence from independent test sets is necessary before an algorithm can be safely translated into clinical practice, especially if its application aims to be generalized in multiple clinical centers (Cearns, Hahn et al. 2019). In addition, some of these algorithms are based on predictors that may represent a significant barrier to their clinical adoption due to their high costs or non-routine assessment in current clinical practice (Hoexter, Miguel et al. 2013, Yun, Jang et al. 2015, Mas, Gasso et al. 2016, Lenhard, Sauer et al. 2018, Reggente, Moody et al. 2018). Besides, two of them are focused on very peculiar treatments or OCD populations (Lenhard, Sauer et al. 2018, Metin, Balli Altuglu et al. 2020). Nevertheless, three studies showed promising predictive performances using only information easy to be assessed in clinical practice (Salomoni, Grassi et al. 2009, Askland, Garnaat et al. 2015, Agne, Tisott et al. 2020),

demonstrating the feasibility of developing clinically translatable ML algorithm for the prediction of OCD clinical course and treatment response prediction.
The present study aims to develop and test a ML algorithm for the prediction of OCD remission after 2 years. To facilitate clinical adoption, only predictors that are easily accessible in the daily clinical routine, such as anamnestic information and questionnaires, were used. The present article reports the results of the first phase with a focus on the preliminary investigation of the generalized predictive performance of the algorithm when applied to new different clinical centers.

Methods

Subjects

Both the training and testing of the algorithm have been performed using data from the Netherlands Obsessive Compulsive Disorder Association (NOCDA) study, a large multi-center naturalistic cohort study of the biological, psychological, and social determinants of chronicity in a clinical sample (Schuurmans, van Balkom et al. 2012). All subjects recruited in the NOCDA study are patients with a lifetime diagnosis of OCD which referred to one of the participating mental health care centers for evaluation and treatment. No formal exclusion criteria were applied except for an inadequate understanding of the Dutch language. The study was approved by the local ethics committees, and all participants gave written informed consent. More details about the rationale, objectives, and methods of NOCDA can be found elsewhere (Schuurmans, van Balkom et al. 2012).
The present study included all NOCDA participants who fulfilled DSM-IV-TR criteria for OCD either at the baseline or at the 2-year follow-up assessment, and whose diagnostic status was respectively reassessed at the 2-year and 4-year follow-up (n= 287). The latter reassessment was used as a 2-year outcome the algorithm aims to predict. In case a subject took part in all baseline, 2-year, and 4-year assessments and fulfilled diagnostic criteria for OCD both at baseline and the 2-year follow-up, it

was included twice in the analyses. Thus, a total of 462 observations were used in the study.
Remission was defined as an absence of the previously present diagnosis of OCD according to the DSM-IV-TR OCD criteria, assessed by the Structured Clinical Interview for the DSM-IV-TR (SCID-I/P) (First, Spitzer et al. 2002), as suggested by international expert consensus (Mataix-Cols, Fernandez de la Cruz et al. 2016). One-hundred and eleven (n=111, 24.03%) remissions were observed.
The subjects have been recruited from eight different clinical centers. Almost half of the sample has been recruited in one center (center Tr: subjects = 131/45.64%, observation = 213/46.10%) and the remaining part from the seven other ones, with a large variation in their contribution, ranging from 10 to 53 subjects and 15 to 87 observations. A detailed description of the number of subjects recruited in each center and the distribution of the remission variable can be found in Table 1.

Features

A detailed description of the information assessed in the NOCDA study is available in the paper addressing the design and rationale of the study (Schuurmans, van Balkom et al. 2012). Only variables available both at baseline and at 2-year assessment were included in the present study. Genetic and biomarker-based variables were discarded because this study aimed to use only information collectible in a clinical interview and with psychometric scales. Two additional variables were defined: current use of a serotoninergic antidepressant and current pharmacological treatment according to the clinical guidelines (Balkom, Vliet et al. 2013).
Some of the variables were not available for all observations and it was a priori decided to remove variables with greater than 20% missing values in the train set (i.e., data coming from the center Tr). Moreover, we included only the categorical predictors in which at least two of the classes had a frequency of at least 5% in the training set, excluding missing values. This was applied to avoid the inclusion of categorical variables whose variation was too small in the training sample. All variables initially included as predictors during the training of the algorithm are reported in Appendix III as supplementary materials.

Two hundred twenty-seven (n=227) features were initially considered. Continuous variables were standardized (mean = 0, standard deviation = 1). Categorical variables were re-coded with the so-called label-encoding strategy, i.e., all cases of each categorical variable have been assigned to an integer number starting from 0. If the variable was ordinal, the class-to-integer conversion respected the order of the classes. In case a "Not answered" class was present, this was not coded as a missing value but the value 0 and other classes starting from 1, because the "Not answered" class may give an additional piece of information rather than a pure missing value (i.e., the subject decided to actively decline to answer instead of that the answer was not collected). The encoding was performed using only the classes occurring in the data used for training. The test data were coded following the coding scheme used for the training data. Any additional class that occurred only in the test dataset was coded as a missing value. This coding strategy for categorical variables is justified by the use of a tree-based ML technique. Missing values were imputed using the MissForest technique (Stekhoven and Bühlmann 2011), implemented with the *IterativeImputer* function of the Scikit-Learn library version 0.22.2 (Pedregosa, Varoquaux et al. 2011) and using Random Forest (Breiman 2001) as estimator. The imputation model was first trained using only the train set and then applied also to the test set.

Gradient boosting technique

Boosting is a ML technique that produces a prediction model in the form of an ensemble of several simpler and consecutively developed prediction models, which are expected to show weaker predictive performance if applied singularly. In our study, we used decision tree models, which is the most common choice within the gradient boosting ensemble technique. Several decision trees are iteratively built, each one consecutively trained to better predict the cases misclassified by the previous model or, as in the case of the gradient boosting approach we used in this study, to predict the error in the prediction performed by the previous model (Friedman 2001). In the end, the final prediction is the result of a weighted sum of the prediction performed by all weaker (up to hundreds) models.

The present study used the implementation of gradient boosted decision trees (GBDT) provided in the eXtreme Gradient Boosting (XGBoost) library (Chen and Guestrin 2016), which is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. This library implements several advancements compared to the standard GBDT technique, among which the possibility of adding stochasticity (Blagus and Lusa 2015) and the use of parallel decision trees (bagging) in each bagging iteration.

Hyper-parameter optimization

As for most of the ML techniques, several hyper-parameters are available for XGBoost, which allow a different tuning of the algorithm during the training process. Different values of these hyper-parameters lead to different predictive performances. The aim is to identify the configuration that produces the best possible performance when applied to cases that are not part of the training set. In order to optimize such hyper-parameters, the algorithm was first trained with 50 random hyperparameter configurations. Subsequently, 150 further configurations were progressively estimated with a Bayesian optimization approach. Bayesian optimization aims to estimate the hyper-parameter configuration that maximizes the performance of the algorithm starting from the previous estimates. It is based on the assumption of a relationship between the various hyper-parameter values and the performance achieved by the algorithm. Bayesian optimization is expected to identify better hyper-parameter configurations with fewer attempts, compared to a random generation of configurations. Estimation was performed with Gaussian Processes, as implemented in the Scikit-Optimized Python library (https://scikit-optimize.github.io/).
The Area Under the Receiving Operating Curve (AUROC) was used as the performance metric to be maximized. The algorithm outputs a continuous prediction score (range: 0-1; the closer to 1 the higher the predicted probability of remission for that subject). The AUROC value can be interpreted as the probability that a randomly selected remitted subject will receive a higher output score than a randomly selected non-remitted subject. The AUROC value is 0.5 when the algorithm makes

random predictions and 1 in case it is always correct in making predictions. AUROC is not affected by class imbalance, and it is independent from any specific threshold that is applied to perform a dichotomous prediction.

Cross-validation

The aim is to train an algorithm that achieves the best possible generalized performance and that also performs well beyond the cases used in the training process. Cross-validation provides an estimate of such generalized performance for every hyper-parameter configuration. In cross-validation, the training sample is divided into several folds of cases that are held-out from the training process, with training iteratively performed with the remaining cases. After the training, the algorithm is finally applied to the held-out cases.
In this study, the commonly used 10-fold cross-validation procedure was applied. The fold creation was performed at random, stratifying (i.e., balancing) for the percentage of remitters and non-remitters in each fold. Finally, the 10 performance estimates of the algorithm available for each hyper-parameter configuration were averaged to provide a final point estimate of the generalized performance. The hyper-parameter configuration that demonstrated the best average cross-validated AUROC was retained and used to retrain a single algorithm with the entire train sample.

Train/test protocol

For training and cross-validation of the final algorithm, all observations from the center Tr were used. The observations from the other seven centers (centers A-G) were used as an independent test set to investigate the predictive performance of the algorithm. Even if sometimes two observations from the same subjects have been included in the analysis (i.e., the baseline assessment information as predictors and the OCD diagnosis at the two-year follow-up as outcome, and the two-year follow-up assessment information as predictors and the OCD diagnosis at the four-year follow-up as outcome), the entire train and test sets are fully independent with respect to the subjects because the test set

(Center A-G) includes observations from subjects that are distinct from those included in the training set (Center Tr).
The algorithm initially outputs a continuous prediction to which a threshold is applied to obtain the final dichotomous prediction of remission. Different threshold values may result in different predictive performances in terms of sensitivity and specificity. In this preliminary investigation, we chose the threshold value which maximizes the balanced accuracy (i.e., the average between sensitivity and specificity) of the cross-validated predictions in the training dataset. This value was applied to obtain the final prediction in the test dataset.
In every single center of the test set, the achieved AUROC, balance accuracy, sensitivity (i.e., Recall), specificity, positive predictive value (i.e., Precision), and negative predictive value were calculated separately. The 95% confidence intervals (CIs) were calculated with a stratified bootstrap procedure, with 10000 resamples (Efron 1987). Only five of the seven centers were considered in these analyses given that two centers provided a small number of observations (Center E = 19; Center F = 15), in which the observed cases of remissions were very limited (Center E = 1; Center F = 3).
The bootstrap resampling technique was also used to investigate if the differences observed in the predictive performance between the different centers were statistically significant. For each statistic, we generated a stratified bootstrap distribution (10000 resamples) of the pairwise differences between two centers and subsequently calculated CIs of the differences, using the very conservative 99.5% range in order to correct for the 10 pairwise comparisons for each statistic (alpha = 0.05/10 = 0.005). A difference was considered statistically significant if both bounds of the CI being above or below the value 0.

Feature importance

Some of the predictors initially taken into consideration may be automatically discarded during the training process. As the NOCDA dataset includes a very extensive assessment, this step may help to reduce the amount of necessary information.
At first, we investigated which predictors were included in the final model. Subsequently, we ranked the retained predictors by importance

using the gain feature importance metric as provided by the XGBoost library. The gain metric indicates the relative contribution of a feature to the model, which is calculated by taking into account the improvement in accuracy brought by that feature at each node split in the ensemble of decision trees (https://xgboost.readthedocs.io/en/release_1.5.0/R-package/discoverYourData.html).

Both the inclusion of a predictor in the model as well as its gain importance score cannot be considered as an absolute metric of the strength of association between the predictor and the probability of the 2-year remission. The inclusion as well as the gain score are closely related to the contribution that a certain predictor has in improving the predictive performance of the specific algorithm that has been developed. This contribution may substantially vary when using other ML techniques, or even with the same technique but with a different hyper-parameter configuration.

Results

Descriptive statistics of all baseline assessment variables are available in Appendix III as supplementary materials, separately for the training and test dataset. Statistics of continuous features are reported before the standardization was applied. In particular, in the train dataset (center Tr), the recruited subjects had a mean age of 39.95 years (SD = 10.75), a mean Y-BOCS total severity compulsions score of 10.26 (SD= 4.28), and a mean Y-BOCS total severity obsessions score of 9.95 (SD = 3.83). In the test dataset (center A-E), the recruited subjects had a mean age of 36.47 years (SD = 11.02), a mean Y-BOCS total severity compulsions score of 10.33 (SD= 4.28), and a mean Y-BOCS total severity obsessions score of 10.6 (SD = 4.09). In the train dataset 118 (55%) were female, while in the test dataset 130 (52.21%) were female.

Performance of the predictive algorithm

The hyper-parameter optimization identified the best hyper-parameter configuration[8] that resulted in an average cross-validated AUROC of 0.7392. The cross-validated predictions obtained with this configuration were pooled together and used to identify the cut-off threshold that maximized the cross-validated balanced accuracy. The obtained threshold value was 0.2193. Applying this threshold to the cross-validated predictions, a balanced accuracy of 71.90%, a sensitivity of 80.00%, a specificity of 68.71%, a positive predictive value of 40.40%, and a negative predictive value of 91.23% were observed. This hyper-parameter configuration was subsequently used to train the final model using the entire train set without cross-validation.

When the final model was tested using the data collected in the centers A, B, C, D, and G, the average AUROC among the centers resulted 0.7820 (95% bootstrap CI = 0.7119-0.8267). Considering the categorical predictions generated with the threshold identified above, results indicated an average balanced accuracy of 72.44% (95% bootstrap CI = 66.81%-77.73%), an average sensitivity of 73.42% (95% bootstrap CI = 65.84%-82.91%), an average specificity of 71.45% (95% bootstrap CI = 63.27%-76.74%), an average positive predictive value of 48.52% (95% bootstrap CI = 40.76%-54.75%), and an average negative predictive value of 87.33% (95% bootstrap CI = 83.94%-92.12%).

When testing the distinct predictive performance of the algorithm per center, results demonstrated a large between-center variation with the AUROC ranging from 0.6364 (A) to 0.9063 (D), the balanced accuracy from 58.02% (A) to 87.50% (D), the sensitivity from 45.45% (A) to 100% (D), the specificity from 62.69% (C) to 76.92% (G), the positive predictive value from 31.25% (A) to 78.57% (G), and the negative predictive value from 78.95% (B) to 100% (D). All point estimates and the 95% bootstrap CIs of the results per center are summarized in Table 3.

Bootstrap analyses revealed significantly different balanced accuracies between center A and center D (58.02% versus 87.50%) and between

8 The resulted best hyperparameter configuration is: base_score = 0.5, booster = 'gbtree', colsample_bylevel = 0.42115547404634657, colsample_bynode = 0.3067377514618746, colsample_bytree = 0.4082812237129432, gamma = 0.9, learning_rate = 0.3, max_delta_step = 1, max_depth = 2, min_child_weight = 0.99, n_estimators = 231, num_parallel_tree = 10, reg_alpha = 0.11711395279718309, reg_lambda = 15.276374168654078, subsample = 0.2.

center C and center D (66.34% versus 87.50%), significantly different sensitivities between center A and center D (45.45% versus 100%) and between center C and center D (70.00% versus 100%), significantly different positive predictive values between center A and center G (31.25% versus 78.57%) and between center C and center G (35.90% versus 78.57%), and significantly different negative predictive values between center A and center D (79.31% versus 100.00%) and between center C and center D (87.50% versus 100.00%). Bootstrap median and 99.5% CIs of the differences are reported in Table 4. In summary, despite the conservative multiple-comparison correction applied in these analyses, the performance of the algorithm sometimes differs considerably between different clinical centers, even though all centers followed the assessment protocol as demanded by the NOCDA study.

Feature importance

The final model included 217 out of the 227 initial features (95.59%), while only 10 variables (4.41%) were discarded. A detailed description of the retained variables, the associated gain feature importance score, and the ranking are reported in Appendix III as supplementary materials.
Based on the gain feature importance metric, the variables ranked as the ten most important predictors in the present algorithm are (Table 2): the total score Y-BOCS severity (Goodman, Price et al. 1989); hours spent every week by the respondent as an organizer of social organizations and clubs (e.g., employers, religious, sport, political or patients organizations); the use of antidepressant drugs on doctors order in the last two weeks; whether the respondent had a paid job at the moment of the baseline assessment; chronic course of OCD in the last two years; the use of any psychotropic drug on doctors order in the last two weeks; participation in sports clubs; the use of psychoanaleptic drugs on doctors order in the last two weeks (defined according to the ATC classification; Organization 2011); the number of different psychotropic drugs currently taken by the subject (defined according to the ATC classification; Organization 2011); and the number of hours the subject work in a week.

Discussion

The present study aimed to develop and test a preliminary ML algorithm for the prediction of the two-year remission in subjects with OCD using data from a large naturalistic multi-center study (NOCDA). Solely predictors based on information from clinical interviews and psychometric scales were included as features. The algorithm was developed and trained using a large sample of subjects recruited in a single center, which represented almost half of the entire dataset. Subsequently, the algorithm was tested in the other participating NOCDA centers. This was done to mimic the translation from a research environment into clinical practice, where new algorithms or protocols are commonly developed in one large center and subsequently applied in smaller centers.

The strict separation between the training and the test set was chosen to increase independence between both datasets. It ensures a sound testing of the generalized performance of the algorithm when applied to clinical centers distinct from the training center.

In this preliminary phase, we arbitrary decided to give equal importance to sensitivity and specificity by defining the predictive threshold that maximized the balanced accuracy in the training dataset. Results showed a moderate predictive performance, with a similar cross-validated and average test balanced accuracy of respectively 71.90% and 72.44%. There is one previous study (Askland, Garnaat et al. 2015) which also aimed to develop a ML model to predict OCD remission based on features assessed by an extensive clinical interview and several psychometric questionnaires. They reported an unbalanced accuracy of 75.4% as the performance of their algorithm. Although there are similarities in the study designs (e.g., both studies are large naturalistic multi-center follow-up studies), this study is not fully comparable to ours because of the performance metrics Askland and colleagues used (e.g., unbalanced accuracy in their study, and balanced accuracy in the current study), and a different definition of OCD remission (at least one period of eight consecutive weeks of sub-threshold symptoms during the entire study enrollment, versus lack of fulfillment of DSM-IV-TR criteria for OCD at the 2-year follow-up assessment in the current study).

The other performance statistics also resulted somewhat similar between cross-validation and testing, with a partial reduction in the average sensitivity and average negative predictive value, and partial improvement in the average specificity and average positive predictive value in the test dataset compared to the cross-validated results obtained in the training dataset. Thus, when the average test performance is taken into account, it might be concluded that the algorithm maintained its performance levels when applied to new clinical centers.

However, in subsequent testing using every single center as a distinct test data set, a substantial variation in the performance statistics was observed between the five centers. The predictive performance of the algorithm was particularly good in some of them, while quite reduced and poor in others. Some between-center differences resulted statistically significant even after a conservative Bonferroni correction for multiple comparisons. Based on these results, any expected performance cannot be guaranteed when the current version of the algorithm is applied to new clinical settings.

Differences in remission rates between the centers (varying from 14.3% to 48%) may affect the predictive performance, but it does not sufficiently explain all of the variability, because also statistics that are in theory unaffected by the rate of remissions occurring in a specific center showed this variation, such as the balanced accuracy, sensitivity, and specificity. The distribution of the characteristics of the OCD subjects may also affect the predictive performance. OCD often is described as a heterogeneous disorder (Mataix-Cols, Rosario-Campos et al. 2005), and also the participants of the NOCDA study were a diverse group. OCD patients referred to a certain clinical center might differ significantly from those referred to another center. The predictive accuracy of a ML algorithm is not necessarily constant among subjects with different characteristics, and some centers may present a higher prevalence of subjects in which the algorithm tends to be less accurate in its predictions. Moreover, variations in the distribution of the predictors (i.e., covariate shift [Shimodaira 2000]) or of the outcome variable (i.e., label shift [Lipton, Wang et al. 2018]) are known to potentially affect the performance of ML algorithms. Besides, even a change in the relationship between the predictor and outcome variables (i.e., concept

drift [Gama, Žliobaitė et al. 2014]) can occur over time and among different populations. Thus, before a medical predictive model can be safely applied in clinical practice, it is crucial to test it not only in a single but in multiple datasets that are independent both to each other and to the data used during the development of the algorithm. As a matter of facts, the majority of medical device filings to regulatory bodies such as the US Food and Drugs Administration are based on multi-center clinical studies (Johnston, Dhruva et al. 2020), and multi-centric testing seems to have progressively become more and more used in the recent literature of ML for medical applications (Abraham, Milham et al. 2017, Meyer, Mueller et al. 2017, Gabr, Coronado et al. 2019).

However, previous studies using ML to predict clinical course and treatment response prediction in OCD patients are mostly based on data recruited in a single center (Salomoni, Grassi et al. 2009, Hoexter, Miguel et al. 2013, Yun, Jang et al. 2015, Mas, Gasso et al. 2016, Lenhard, Sauer et al. 2018, Reggente, Moody et al. 2018, Metin, Balli Altuglu et al. 2020). Only Askland and colleagues (Askland, Garnaat et al. 2015) used a large multi-center dataset from a longitudinal study of OCD (The Brown Longitudinal Obsessive-Compulsive Study Pinto; Mancebo et al. 2006). However, pooled data from all centers were used both for training and testing. Their testing was not designed to ensure center-independence from the data used during the training of the algorithm and thus the predictive performance may differ when the algorithm is applied to new, independent data sets. In conclusion, the present study is the first one using ML in OCD course prediction which tested the algorithm in an independent test sample consisting of data from other centers than the training center.

Some strategies that attempt to reduce the impact of the above-mentioned distribution shift/drift have been proposed in the ML literature (Shimodaira 2000, Gama, Žliobaitė et al. 2014, Lipton, Wang et al. 2018). However, any application of such correction strategies requires advanced knowledge of the predictor and/or target variable distributions in the particular setting where the algorithms will be used. Thus, a relevant amount of data has to be preliminary available for any new center, or these data have to be collected in advance for the sole purpose of developing the center-specific correction of the algorithm. Especially for the outcome variable, which is based on a 2-year follow-

up, this preliminary data collection would be particularly burdensome and may delay the introduction of the algorithm in a particular clinical center.

Another potential strategy to reduce the impact of variable distribution shifts/drifts is to include only predictive variables in the algorithm with more stable distributions among clinical centers, and a stable relationship with the outcome variable. A reduction of the number of predictors may also help to improve the applicability of the algorithm in the daily clinical practice. Although the present algorithm only uses information from clinical interviews and questionnaires, the extensive NOCDA assessment protocol is time-consuming and may be exhausting for patients. Unfortunately, less than 5% of the features were automatically discarded during the training process, which is a characteristic of the GBDT technique, and the algorithm still relies on 217 predictors. A further reduction of the predictive variables will be later performed by applying some additional feature selection strategies, by taking into account the gain feature importance metric, and by evaluating the clinical importance and availability of the predictors. This may lead to the development of a more robust algorithm while maintaining or perhaps even improving its predictive performance.

The ranking of importance of the predictors based on the gain feature importance metric confirms that factors of different nature may contribute to the prognosis of OCD, without one domain being the sole or primary source of it. For example, the ten features that resulted as most relevant are related to very different domains, such as clinical severity and characterizations, medications, work, and social activities. This also supports the necessity of using models that simultaneously consider multiple predictors rather than individual factors to achieve relevant prediction of OCD remission and prognosis.

Some limitations should be taken into account. We used GBDT as the sole ML technique in our study. Several other supervised ML techniques exist, all of which may have led to different results. Some of these other techniques may have even resulted in better predictive performance, and an ensemble of different techniques can also be used in the attempt of achieving better results (Grassi, Rouleaux et al. 2019). In this preliminary phase, we opted to focus on the GBDT technique for several reasons. First, it has proved to be a powerful technique even if used individually

(Natekin and Knoll 2013). Moreover, given the large number of categorical variables we used as predictors, this technique was chosen because it can handle non-dichotomous categorical variables with efficient coding strategies (e.g., label encoding), allowing the use of a single predictor per categorical variable instead of a single predictor per class of each categorical variable (i.e., one-hot encoding), as it is required by most of the other supervised ML techniques. Furthermore, a metric of the importance of the predictive variables, i.e., the gain feature importance metric, can be derived natively and computationally efficiently directly from the algorithm, which takes into account the interactions between the predictive variables and does not require additional analyses to be performed after the final model has been trained. Finally, less important features are expected to be discarded automatically during the development of the algorithm, with the GBDT technique operating an automatic and model-tailored feature selection. Considering all these characteristics of the GBDT technique, it seemed convenient to use this single technique in this preliminary step, leaving the use of further techniques and their ensembling to the following phases of our research.

Another limitation is that, although independent to each other, the clinical centers of the NOCDA study all collected the data following the same assessment protocol. Moreover, they are all located in the Netherlands. Thus, these centers may share more similarities than other clinical centers not following the standardized assessment protocol or from other countries. Therefore, when the algorithm is applied to new clinical centers, the predictive performance may vary even more compared to the variation observed in the present study.

An additional limitation is that, although two centers with a very limited number of cases were excluded, the sample size of the data from some of the five test centers was small. In the next phase, we plan to include the final follow-up assessment (predictor variables from the four-year follow-up assessment and remission at the six-year follow-up assessment) to enlarge the sample size for both in training and testing of the algorithm.

Finally, the definition of remission used in this study is the absence of an OCD diagnosis at the two-year follow-up assessment. As the course is not unidirectional but shows periods of remission and subsequent

relapse in the majority of the OCD patients (Reddy, D'Souza et al. 2005, Eisen, Sibrava et al. 2013, Garnaat, Boisseau et al. 2015), the current prediction of the algorithm may not be able to provide an exhaustive description of the clinical course of the subject. A more complex modeling of the course of OCD may be desirable, based on information about the course of OCD assessed longitudinally during several follow-ups.

Some strengths may also be mentioned. Data are based on a large naturalistic multi-center study. The longitudinal design, with baseline and successive follow-up assessments, makes the application of ML techniques particularly suitable to examine predictors of the course of OCD. The naturalistic investigation of the illness course contributes to the clinical validity of the ML algorithms developed with data from the NOCDA study. All features can be assessed during the daily routing using interviews and questionnaires, which makes it easily accessible. With a total of 462 observations, it is one of the largest ML studies in the field of OCD research. Approximately half of the subjects were recruited from a single center, and the remaining part of the sample from the other seven centers, with a large variation in the number of subjects recruited in each one of them. This mimics the common scenario in which a larger dataset coming from one or a few centers is used to train a ML algorithm, which will be later applied to other centers. In contrast to previous studies in this field, the present study did not only develop an algorithm for OCD course prediction and tested it within the training set, but also applied a thorough testing by subsequently validating the algorithm in a test sample consisting of data from other centers than the training center.

The present study aimed to develop a clinically accessible algorithm that predicts remission of OCD, which is based on information that can be easily assessed in the daily clinical routine. However, if this information is not sufficient to achieve a good level of prediction, the inclusion of additional predictors, such as genetic or neuroimaging biomarkers, should be investigated. Although costs and availability can make their introduction in the clinical routine quite challenging, it may be justified if they significantly increase the predictive performance of the algorithm given the contribution that a prediction of the OCD course may bring to treatment planning and appropriate support of OCD patients. In the

NOCDA study, further biological and genetic information has been collected and we also plan as a further next step to investigate if the addition of such information may relevantly increase the accuracy of our algorithm.
In conclusion, the present study developed and tested a ML algorithm for the prediction of the 2-year remission of the diagnosis of OCD using data from a large, multi-center study (NOCDA). The algorithm was development with data coming from one large clinical center and subsequently tested with data from different smaller centers. Results evidenced a moderate average generalized performance but showed a large variation between the centers when investigated per distinct center. This demonstrates the difficulties algorithms have to overcome before they can be safely translated from the research environment into clinical practice. It also emphasizes the need for independent test samples from different centers during further research.

References

Abraham, A., M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion and G. Varoquaux (2017). "Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example." NeuroImage 147: 736-745.

Agne, N. A., C. G. Tisott, P. Ballester, I. C. Passos and Y. A. Ferrão (2020). "Predictors of suicide attempt in patients with obsessive-compulsive disorder: an exploratory study with machine learning analysis." Psychological Medicine: 1-11.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders, 5th. Washington, DC.

Askland, K. D., S. Garnaat, N. J. Sibrava, C. L. Boisseau, D. Strong, M. Mancebo, B. Greenberg, S. Rasmussen and J. Eisen (2015). "Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy." International journal of methods in psychiatric research **24**(2): 156-169.

Balkom, A. v., I. v. Vliet, P. Emmelkamp, C. Bockting, J. Spijker and n. d. W. M. r. A. D. (2013) (2013). Multidisciplinaire richtlijn

Angststoornissen (Derde revisie). Richtlijn voor de diagnostiek, behandeling en begeleiding van volwassen patiënten met een angststoornis. Utrecht, Trimbos-instituut.

Blagus, R. and L. Lusa (2015). "Boosting for high-dimensional two-class prediction." BMC Bioinformatics **16**: 300.

Breiman, L. (2001). "Random Forests." Machine Learning **45**(1): 5-32.

Cearns, M., T. Hahn and B. T. Baune (2019). "Recommendations and future directions for supervised machine learning in psychiatry." Transl Psychiatry **9**(1): 271.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Cherian, A. V., S. B. Math, T. Kandavel and Y. C. Reddy (2014). "A 5-year prospective follow-up study of patients with obsessive-compulsive disorder treated with serotonin reuptake inhibitors." J Affect Disord **152-154**: 387-394.

Coluccia, A., A. Fagiolini, F. Ferretti, A. Pozza, G. Costoloni, S. Bolognesi and A. Goracci (2016). "Adult obsessive-compulsive disorder and quality of life outcomes: A systematic review and meta-analysis." Asian J Psychiatr **22**: 41-52.

Efron, B. (1987). "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association **82**(397): 171-185.

Eisen, J. L., N. J. Sibrava, C. L. Boisseau, M. C. Mancebo, R. L. Stout, A. Pinto and S. A. Rasmussen (2013). "Five-year course of obsessive-compulsive disorder: predictors of remission and relapse." J Clin Psychiatry **74**(3): 233-239.

Ezzati, M., A. D. Lopez, A. A. Rodgers and C. J. Murray (2004). Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors, World Health Organization.

Fineberg, N. A., A. Brown, S. Reghunandanan and I. Pampaloni (2012). "Evidence-based pharmacotherapy of obsessive-compulsive disorder." Int J Neuropsychopharmacol **15**(8): 1173-1191.

First, M. B., R. L. Spitzer, M. Gibbon and J. Williams (2002). Structured clinical interview for DSM-IV-TR Axis I disorders, research version.

Friedman, J. H. (2001). "Greedy function approximation: a gradient boosting machine." Annals of statistics: 1189-1232.

Gabr, R. E., I. Coronado, M. Robinson, S. J. Sujit, S. Datta, X. Sun, W. J. Allen, F. D. Lublin, J. S. Wolinsky and P. A. Narayana (2019). "Brain and lesion segmentation in multiple sclerosis using fully convolutional neural networks: A large-scale study." Mult Scler: 1352458519856843.

Gama, J., I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia (2014). "A survey on concept drift adaptation." ACM computing surveys (CSUR) **46**(4): 1-37.

Garnaat, S. L., C. L. Boisseau, A. Yip, N. J. Sibrava, B. D. Greenberg, M. C. Mancebo, N. C. McLaughlin, J. L. Eisen and S. A. Rasmussen (2015). "Predicting course of illness in patients with severe obsessive-compulsive disorder." J Clin Psychiatry **76**(12): e1605-1610.

Goodman, W. K., L. H. Price, S. A. Rasmussen, C. Mazure, P. Delgado, G. R. Heninger and D. S. Charney (1989). "The Yale-Brown Obsessive Compulsive Scale. II. Validity." Arch Gen Psychiatry **46**(11): 1012-1016.

Grassi, M., N. Rouleaux, D. Caldirola, D. Loewenstein, K. Schruers, G. Perna and M. Dumontier (2019). "A Novel Ensemble-Based Machine Learning Algorithm to Predict the Conversion From Mild Cognitive Impairment to Alzheimer's Disease Using Socio-Demographic Characteristics, Clinical Information, and Neuropsychological Measures." Front Neurol **10**: 756.

Hazari, N., J. C. Narayanaswamy and S. S. Arumugham (2016). "Predictors of response to serotonin reuptake inhibitors in obsessive-compulsive disorder." Expert Rev Neurother **16**(10): 1175-1191.

Hoexter, M. Q., E. C. Miguel, J. B. Diniz, R. G. Shavitt, G. F. Busatto and J. R. Sato (2013). "Predicting obsessive-compulsive disorder severity combining neuroimaging and machine learning methods." J Affect Disord **150**(3): 1213-1216.

Johnston, J. L., S. S. Dhruva, J. S. Ross and V. K. Rathi (2020). "Clinical Evidence Supporting FDA Clearance of First-of-a-Kind Therapeutic Devices via the De Novo Pathway Between 2011 and 2019." medRxiv: 2020.2004.2023.20077164.

Keeley, M. L., E. A. Storch, L. J. Merlo and G. R. Geffken (2008). "Clinical predictors of response to cognitive-behavioral therapy for obsessive-compulsive disorder." Clin Psychol Rev **28**(1): 118-130.

Kempe, P. T., P. van Oppen, E. de Haan, J. W. Twisk, A. Sluis, J. H. Smit, R. van Dyck and A. J. van Balkom (2007). "Predictors of course in obsessive-compulsive disorder: logistic regression versus Cox regression for recurrent events." Acta Psychiatr Scand **116**(3): 201-210.

Knopp, J., S. Knowles, P. Bee, K. Lovell and P. Bower (2013). "A systematic review of predictors and moderators of response to psychological therapies in OCD: Do we have enough empirical evidence to target treatment?" Clinical Psychology Review **33**(8): 1067-1081.

Lenhard, F., S. Sauer, E. Andersson, K. N. Mansson, D. Mataix-Cols, C. Ruck and E. Serlachius (2018). "Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach." Int J Methods Psychiatr Res **27**(1).

Lipton, Z., Y.-X. Wang and A. Smola (2018). Detecting and Correcting for Label Shift with Black Box Predictors. International Conference on Machine Learning.

Mas, S., P. Gasso, A. Morer, A. Calvo, N. Bargallo, A. Lafuente and L. Lazaro (2016). "Integrating Genetic, Neuropsychological and Neuroimaging Data to Model Early-Onset Obsessive Compulsive Disorder Severity." PLoS One **11**(4): e0153846.

Mataix-Cols, D., L. Fernandez de la Cruz, A. E. Nordsletten, F. Lenhard, K. Isomura and H. B. Simpson (2016). "Towards an international expert consensus for defining treatment response, remission, recovery and relapse in obsessive-compulsive disorder." World Psychiatry **15**(1): 80-81.

Mataix-Cols, D., M. C. Rosario-Campos and J. F. Leckman (2005). "A multidimensional model of obsessive-compulsive disorder." Am J Psychiatry **162**(2): 228-238.

Metin, S. Z., T. Balli Altuglu, B. Metin, T. T. Erguzel, S. Yigit, M. K. Arıkan and K. N. Tarhan (2020). "Use of EEG for Predicting Treatment Response to Transcranial Magnetic Stimulation in Obsessive Compulsive Disorder." Clinical EEG and Neuroscience **51**(3): 139-145.

Meyer, S., K. Mueller, K. Stuke, S. Bisenius, J. Diehl-Schmid, F. Jessen, J. Kassubek, J. Kornhuber, A. C. Ludolph, J. Prudlo, A. Schneider, K. Schuemberg, I. Yakushev, M. Otto and M. L. Schroeter (2017). "Predicting behavioral variant frontotemporal dementia with pattern classification in multi-center structural MRI data." Neuroimage Clin **14**: 656-662.

Nakajima, A., N. Matsuura, K. Mukai, K. Yamanishi, H. Yamada, K. Maebayashi, K. Hayashida and H. Matsunaga (2018). "Ten-year follow-up study of Japanese patients with obsessive-compulsive disorder." Psychiatry Clin Neurosci **72**(7): 502-512.

Natekin, A. and A. Knoll (2013). "Gradient boosting machines, a tutorial." Front Neurorobot **7**: 21.

Organization, W. H. (2011). "WHO collaborating centre for drug statistics methodology. ATC/DDD index 2011." World Health Organization2011WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD index.

Ost, L. G., A. Havnen, B. Hansen and G. Kvale (2015). "Cognitive behavioral treatments of obsessive-compulsive disorder. A systematic review and meta-analysis of studies published 1993-2014." Clin Psychol Rev **40**: 156-169.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg (2011). "Scikit-learn: Machine learning in Python." the Journal of machine Learning research **12**: 2825-2830.

Pinto, A., M. C. Mancebo, J. L. Eisen, M. E. Pagano and S. A. Rasmussen (2006). "The Brown Longitudinal Obsessive Compulsive Study: clinical features and symptoms of the sample at intake." J Clin Psychiatry **67**(5): 703-711.

Pozza, A., C. Lochner, F. Ferretti, A. Cuomo and A. Coluccia (2018). "Does higher severity really correlate with a worse quality of life in

obsessive-compulsive disorder? A meta-regression." Neuropsychiatr Dis Treat **14**: 1013-1023.

Reddy, Y. C., S. M. D'Souza, C. Shetti, T. Kandavel, S. Deshpande, S. Badamath and S. Singisetti (2005). "An 11- to 13-year follow-up of 75 subjects with obsessive-compulsive disorder." J Clin Psychiatry **66**(6): 744-749.

Reggente, N., T. D. Moody, F. Morfini, C. Sheen, J. Rissman, J. O'Neill and J. D. Feusner (2018). "Multivariate resting-state functional connectivity predicts response to cognitive behavioral therapy in obsessive-compulsive disorder." Proc Natl Acad Sci U S A **115**(9): 2222-2227.

Salomoni, G., M. Grassi, P. Mosini, P. Riva, P. Cavedini and L. Bellodi (2009). "Artificial neural network model for the prediction of obsessive-compulsive disorder treatment response." J Clin Psychopharmacol **29**(4): 343-349.

Samuel, A. L. (1959). "Some studies in machine learning using the game of checkers." IBM Journal of research and development **3**(3): 210-229.

Schuurmans, J., A. J. van Balkom, H. J. van Megen, J. H. Smit, M. Eikelenboom, D. C. Cath, M. Kaarsemaker, D. Oosterbaan, G. J. Hendriks, K. R. Schruers, N. J. van der Wee, G. Glas and P. van Oppen (2012). "The Netherlands Obsessive Compulsive Disorder Association (NOCDA) study: design and rationale of a longitudinal naturalistic study of the course of OCD and clinical characteristics of the sample at baseline." Int J Methods Psychiatr Res **21**(4): 273-285.

Shimodaira, H. (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function." Journal of statistical planning and inference **90**(2): 227-244.

Stekhoven, D. J. and P. Bühlmann (2011). "MissForest—nonparametric missing value imputation for mixed-type data." Bioinformatics **28**(1): 112-118.

van Oppen, P., A. J. van Balkom, E. de Haan and R. van Dyck (2005). "Cognitive therapy and exposure in vivo alone and in combination with fluvoxamine in obsessive-compulsive disorder: a 5-year follow-up." J Clin Psychiatry **66**(11): 1415-1422.

Yun, J. Y., J. H. Jang, S. N. Kim, W. H. Jung and J. S. Kwon (2015). "Neural Correlates of Response to Pharmacotherapy in Obsessive-Compulsive Disorder: Individualized Cortical Morphology-Based Structural Covariance." Prog Neuropsychopharmacol Biol Psychiatry **63**: 126-133.

Table 1. Description of test centers.

	Center	Subjects	Total Observations	Remitters	
				N	%
Included in the testing	Test - A	27	45	11	24.44%
	Test - B	20	30	10	33.33%
	Test - C	53	87	20	22.99%
	Test - D	17	28	4	14.29%
	Test - G	19	25	12	48.00%
	Mean	27	43	11	28.61%
	S.D.	13.33	23.06	5.12	11.42%
	Minimum	17	25	4	14.29%
	Maximum	53	87	20	48.00%
Excluded from the testing	Test - E	10	19	1	5.26%
	Test - F	10	15	3	20.00%

S.D. = Standard Deviation.

Table 2. Description of the ten most important predictors.

Variable	Descriptive Statistics in the Train Set	Descriptive Statistics in the Test Set	Gain Feature Importance (Score)	Gain Feature Importance (Rank)
Total severity score of Y-BOCS	Mean: 19.96, S.D.: 6.94	Mean: 19.9, Yes, S.D.: 7.35	0.0077	1
How many hours a week the respondent is involved in an executive role in a club or organizations (e.g., sport club, music band, organization for patient, social organization, religious organization, political party)?	Mean: 0.17, S.D.: 0.92	Mean: 0.16, S.D.: 0.96.	0.0076	2
Antidepressant use on doctors order in the last two weeks	No: 92 (43%), Yes: 110 (52%), Missing values: 11 (5.16%)	No: 75 (30.12%), Yes: 165 (66.27%), Missing values: 9 (4%)	0.0076	3
Do you have a paid job at the moment?	No, I have never had a paid job: 5 (2%), No, I had a paid job in the past: 71 (33%), Yes: 126 (59%), Missing values: 11 (5.16%)	No, I have never had a paid job: 9 (3.61%), No, I had a paid job in the past: 116 (46.59%), Yes: 115 (46.18%), Missing values: 9 (4%)	0.0075	4
Chronical Course of OCD in the last 2 years	Too many omitted answers from the respondent: 3 (1%), No: 90 (42%), Yes: 120 (56%)	Too many omitted answers from the respondent: 3 (1.2%), No: 97 (38.96%), Yes: 146 (58.63%), Missing values: 3 (1%)	0.0070	5
Psychotropic use on doctors order in the last two weeks	No: 80 (38%), Yes: 122 (57%), Missing values: 11 (5.16%)	No: 66 (26.51%), Yes: 174 (69.88%), Missing values: 9 (4%)	0.0070	6
Dre you involved in sports clubs?	Asked but the respondent did not answer: 5 (2%), No: 134 (63%), Yes: 71 (33%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 165 (66.27%), Yes: 78 (31.33%)	0.0068	7
Psychoanaleptic use on doctors order in the last two weeks	No: 92 (43%), Yes: 110 (52%), Missing values: 11 (5.16%)	No: 72 (28.92%), Yes: 168 (67.47%), Missing values: 9 (4%)	0.0068	8
Total number of different psychotropic medications currently used by the subjects	Mean: 0.99, S.D.: 1.08	Mean: 1.25, S.D.: 1.2	0.0066	9
How many hours did you work a week recently?	Mean: 16.98, S.D.: 16.75	Mean: 12.93, S.D.: 16.37	0.0064	10

The Italic font in the variable description indicates questions asked to the respondent.

Table 3. Test predictive performance.

		Test A	Test B	Test C	Test D	Test G	Mean	S.D.	Minumum	Maximum
Auroc	Point Estimate	0.6364	0.7300	0.7336	0.9063	0.9038	0.782	0.1063	0.6364	0.9063
	95% Bootstrap CI	0.4412	0.5250	0.6187	0.7708	0.7628				
		0.8155	0.9150	0.8381	1	1				
Balanced accuracy	Point Estimate	58.02%	67.50%	66.34%	87.50%	84.29%	72.73%	11.28%	58.02%	87.50%
	95% Bootstrap CI	39.97%	50.00%	54.37%	79.17%	68.59%				
		73.13%	85.00%	77.57%	95.83%	96.15%				
Sensitivity	Point Estimate	45.45%	60.00%	70.00%	100.00%	91.67%	73.42%	20.07%	45.45%	100.00%
	95% Bootstrap CI	18.18%	30.00%	50.00%	100.00%	75.00%				
		72.73%	90.00%	90.00%	100.00%	100.00%				
Specificity	Point Estimate	67.65%	75.00%	62.69%	75.00%	76.92%	71.45%	5.41%	62.69%	76.92%
	95% Bootstrap CI	52.94%	55.00%	50.75%	58.33%	53.85%				
		82.35%	95.00%	74.63%	91.67%	100.00%				
Positive predictive value	Point Estimate	31.25%	54.55%	35.90%	40.00%	78.57%	48.05%	17.14%	31.25%	78.57%
	95% Bootstrap CI	13.33%	33.33%	26.47%	28.57%	62.50%				
		50.00%	81.82%	46.15%	66.67%	100.00%				
Negative predictive value	Point Estimate	79.31%	78.95%	87.50%	100.00%	90.91%	87.33%	7.85%	78.95%	100.00%
	95% Bootstrap CI	70.00%	66.67%	80.00%	100.00%	75.00%				
		89.66%	93.75%	95.35%	100.00%	100.00%				

AUROC = Area under the receiving operating curve. 95% Bootstrap CI (Lower) = lower bound of the 95% bootstrap confidence interval. 95% Bootstrap CI (Higher) = higher bound of the 95% bootstrap confidence interval. S.D. = standard deviation.

Table 4. Pairwise between-centers differences of test predictive performance.

AUROC	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-0.2844	0.4676	/	/	/	/	/	/	/	/
Test - C	-0.2087	0.4298	-0.3035	0.3511	/	/	/	/	/	/
Test - D	-0.0506	0.5875	-0.1533	0.5079	-0.0998	0.3874	/	/	/	/
Test - G	-0.0606	0.5877	-0.1551	0.5062	-0.1005	0.3996	-0.2708	0.2516	/	/

Balanced Accuracy	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-23.58%	45.91%	/	/	/	/	/	/	/	/
Test - C	-19.18%	37.45%	-30.63%	29.59%	/	/	/	/	/	/
Test - D	4.33%	57.10%	-7.50%	48.33%	1.00%	42.15%	/	/	/	/
Test - G	-4.39%	57.31%	-16.60%	48.65%	-9.70%	42.65%	-28.69%	18.75%	/	/

Sensitivity	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-42.73%	72.73%	/	/	/	/	/	/	/	/
Test - C	-26.82%	71.82%	-40.00%	60.00%	/	/	/	/	/	/
Test - D	9.09%	90.91%	0.00%	80.00%	5.00%	60.00%	/	/	/	/
Test - G	-6.06%	90.91%	-16.67%	80.00%	-18.33%	55.00%	-41.67%	/	/	/

Specificity	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-27.65%	41.77%	/	/	/	/	/	/	/	/
Test - C	-31.61%	23.13%	-41.79%	20.67%	/	/	/	/	/	/
Test - D	-26.96%	39.95%	-36.67%	36.67%	-18.66%	40.61%	/	/	/	/
Test - G	-34.39%	45.25%	-41.54%	42.31%	-25.72%	44.78%	-40.71%	41.67%	/	/

Positive Predictive Value	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-19.23%	70.83%	/	/	/	/	/	/	/	/
Test - C	-28.18%	35.39%	-61.77%	17.56%	/	/	/	/	/	/
Test - D	-25.00%	58.98%	-58.93%	36.25%	-18.75%	48.75%	/	/	/	/
Test - G	11.11%	83.33%	-22.79%	65.03%	16.30%	71.05%	-9.41%	69.23%	/	/

Negative Predictive Value	Test - A		Test - B		Test - C		Test - D		Test - G	
	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)	99.5% Bootstrap CI (Lower)	99.5% Bootstrap CI (Higher)
Test - A	/	/	/	/	/	/	/	/	/	/
Test - B	-22.86%	25.69%	/	/	/	/	/	/	/	/
Test - C	-9.95%	25.18%	-14.44%	29.66%	/	/	/	/	/	/
Test - D	5.26%	34.62%	0.00%	38.89%	2.17%	24.00%	/	/	/	/
Test - G	-17.21%	33.33%	-20.77%	36.84%	-23.56%	22.00%	-33.33%	/	/	/

Difference is calculated as center in the row minus center in the column. Statistical significance (i.e., both bounds of the 99.5% boostrap CI higher or lower than 0) evidence in italic. 99.5% Bootstrap CI (Lower) = lower bound of the 99.5% bootstrap confidence interval of the pairwise centers difference. 99.5% Bootstrap CI (Higher) = higher bound of the 99.5% bootstrap confidence interval of the pairwise centers difference.

CHAPTER 6

BETTER AND FASTER AUTOMATIC SLEEP STAGING WITH ARTIFICIAL INTELLIGENCE: A CLINICAL VALIDATION STUDY OF NEW SOFTWARE FOR SLEEP SCORING

Massimiliano Grassi[a,b,c,*], Archie Defillo[c], Daniela Caldirola[a,b], Silvia Daccò[a,b], Koen Schruers[d], Giampaolo Perna[a,b,c,d,e]

[a]Department of Clinical Neurosciences, Hermanas Hospitalarias, Villa San Benedetto Menni Hospital, Albese con Cassano, Como, Italy.
[b]Department of Biomedical Sciences, Humanitas University, Rozzano, Milan, Italy.
[c]Medibio Limited, Savage, Minnesota, USA.
[d]Research Institute of Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine, and Life Sciences, Maastricht University, Maastricht, The Netherlands.
[e]Department of Psychiatry and Behavioral Sciences, Leonard Miller School of Medicine, University of Miami, Florida, USA.

To be submitted.

CHAPTER 7

GENERAL DISCUSSION

APPENDIX I

SUMMARY

As introduced in **Chapter 1**, mental disorders are widespread and burdensome, but most people suffering from them still miss to receive proper treatment or experience treatment unresponsiveness, relapses, and recurrence episodes. Psychiatry is still based on descriptive diagnostic taxonomies with limited validity, and clinical guidelines recommend interventions only for the 'average' patients suffering from a specific diagnostic class. The new paradigm of Personalized Medicine promises to improve the treatment and prevention of mental disorders by providing better individual indications and predictions. Recent technological advancements now permit to cost-effectively collect vast sets of information that may be used to achieve such personalized recommendations. However, data are not enough, and it is also necessary to develop models that can perform these personalized recommendations. Given the complexity and multifactorial nature of mental disorders, it is difficult to fully achieve a full *a priori* understanding of the phenomenon that is then used to develop rules that clinicians are advised to follow during their clinical decision-making process. A promising alternative to develop such tools is the use of Supervised Machine Learning (SML). These techniques use examples in which both the input and the desired output variables are available. From these examples, such techniques are able to automatically extract patterns and build algorithms that can provide an estimate of the output variables in new cases in which only the input variables are known. SML opens the possibility to develop PM tools that may have been impossible otherwise, only by having enough suitable examples and without the need for an explicit *a priori* understanding of the relationship between the input and output variables. Moreover, SML algorithms may help automatize some time-consuming clinical tasks, reducing the associate costs and clinicians' burden. In psychiatric scientific literature, a growing number of research articles using SML are available, and most of them present algorithms that seem to achieve very high performances. However, this amount of promising evidence appears to conflict with the general lack of SML-based tools in psychiatric clinical practice. Several challenges need to be faced to ensure a safe and effective application of these algorithms in clinical practice.

This doctoral dissertation aims to present the development and testing of some SML algorithms for the psychiatric clinical practice, with two main focuses: the ability to achieve good performance by solely using input information that facilitates or at least does not hinder its clinical adoption, and the necessity to provide preliminary evidence of the expected generalized performance of the algorithm even at early stages of its development.

In particular, **Chapters 2**, **3**, and **4** report three studies related to the development of an SML algorithm for the 3-year prediction of conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD). Currently available and emerging therapies for AD likely have the most significant impact when provided at the earliest disease stage. Thus, the possibility to early identify which subjects are at high risk of later developing AD, e.g., subjects with MCI, it is of crucial importance. However, currently proposed machine learning algorithms seem to achieve only limited predictive accuracy, or they are based on expensive and hard-to-collect information.

The study presented in **Chapter 2** aimed to develop an initial proof-of-concept of a clinically-translatable SML algorithm for the 3-years prediction of conversion to AD in MCI and Pre-mild Cognitive Impairment (PreMCI) subjects. This algorithm is based only on non-invasive predictors that are either already routinely assessed or easily introducible in clinical practice. Specifically, baseline information regarding sociodemographic characteristics, clinical and neuropsychological test scores, cardiovascular risk indexes, and a visual rating scale for brain atrophy was used as input. Data were extracted from a longitudinal, multicentric dataset collected in Miami (Florida, US). A subset of 16 predictors was selected from all the abovementioned domains. The best model (support vector machine with radial-basis function kernel) resulted in a high leave-pair-out-cross-validation performance, with an Area Under the Receiver Operating Characteristics (AUROC) of 0.962, a balanced accuracy of 91.3%, a sensitivity of 95.6%, and a specificity of 87.1%. These results are among the best of the many algorithms available in the literature and the best achieved so far using only information easily collectible in clinical practice.

However, these preliminary results are based solely on a cross-validation approach, and not on a set of test examples that have completely been held out during the development of the algorithm. Thus, to provide a sounder estimate of its expected performance, the study in **Chapter 3** aimed to perform indirect testing of this algorithm via a transfer learning approach. The same predictors and SML technique used in the previous study were employed to retrain the algorithm to accomplish the task of discriminating between AD and Cognitively Normal individuals (CN). Data used for training were another sample of subjects with either the former or the latter condition that have being recruited during same longitudinal, multicentric dataset collected in Miami (Florida, US) used in the previous study. The new algorithm was then used to predict the three-year conversion to AD in the same sample of MCI subjects used in the previous study. In this study, the MCI sample was entirely held out during training and only used to test the algorithm. A reduced but still significant predictive performance was observed in the MCI sample (AUROC = 0.821; balanced accuracy = 77.9%; sensitivity = 85.2%; specificity = 70.6%), and these can be considered a first indirect, possibly conservative estimate of the performance of the algorithm presented in **Chapter 2** when applied to a sample of MCI subjects not used in the training process.

Chapter 4 presents an improved algorithm for the same task based on an ensemble of several SML algorithms whose individual predictions are aggregated with a weighted average rank approach. Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) open database were employed in this study. A restricted set of information, which included sociodemographic and clinical characteristics and neuropsychological test scores, was used as predictors, while any imaging information was excluded entirely. Moreover, a peculiar site-independent stratified train/test split protocol was used to better estimate the generalized performance of the algorithm when applied in clinical centers different from those used for training and validation. The ensemble of the SML algorithms demonstrated a test AUROC of 0.88, a sensitivity of 77.7%, and a specificity of 79.9%. In addition, it demonstrated a specificity of 40.2%/53% when the threshold was optimized to achieve a sensitivity of respectively 100% and 95%. These results show evidence of high

predictive accuracy even when testing is performed with a sound train/test split protocol, exhibiting particularly good predictive performance when the algorithm was optimized as a screening tool. Thus, the algorithm may be useful to improve recruitment in clinical trials and to more selectively prescribe newly emerging early interventions to patients at high risk to convert from MCI to AD.

The work described in **Chapter 5** relates to the initial development of an SML algorithm for the prediction of 2-year OCD remission. The OCD course differs widely between OCD patients, varying from severe chronic symptoms to full remission. No tools for individual prediction of OCD remission are currently available. To facilitate clinical adoption, only predictors easily accessible in the daily clinical routine, such as anamnestic information and questionnaires, were used in the algorithm. Gradient boosted decision trees (GBDT) were used as a supervised machine learning technique. The training of the algorithm was performed with 227 features and a sample of 215 cases recruited in a single clinical center. The predictive performance of the algorithm was subsequently tested using an independent sample of 215 cases recruited in five different centers. All data were collected in a longitudinal multi-center study (NOCDA). The predictive performance of the algorithm in the five test centers resulted in an average AUROC of 0.7820, an average balanced accuracy of 72.73%, an average sensitivity of 73.42%, and an average specificity of 71.45%. However, a large between-center variation was observed (AUROC range = 0.636-0.906; balanced accuracy range = 58.0%-87.5%; sensitivity rage = 45.5%-100.0%; specificity range = 62.7%-76.9%), which evidence the challenge of achieving a stable generalized performance when applied into different clinical settings. These results highlight the necessity of testing SML algorithms in samples collected in different sites before being safely translated from the research environment into clinical practice.

Chapter 6 presents an extensive test of a computer program (MEBsleep by Medibio Limited) that performs automatic sleep staging of polysomnography by processing the signals of EEG montages through a processing pipeline of SML algorithms. Sleep staging of polysomnography is a time-consuming task, it requires significant

training, and significant variability among scorers is expected. Testing has been based on the agreement of the staging performed by the program with the manual sleep staging performed by expert sleep technicians. The extensive test performed in this study aim to finally demonstrate the clinical applicability of the SML-based software. Forty polysomnography recordings of patients referred for sleep evaluation to three different sleep clinics were retrospectively collected. Three experienced technicians independently staged the recording twice, first taking into account only the electroencephalography signals, and then also the electromyography and electrooculography signals in compliance with the staging rules recommended by the American Academy of Sleep Medicine guidelines. In addition, the staging performed initially in clinical practice was also considered. Several agreement statistics of the automatic staging with the manual staging, among the different manual staging scoring, and their differences were calculated as a test of the performance achieved by the SML-based application. The automatic staging resulted for the most part comparable or significantly more in agreement with the technicians' staging than the between-technician agreement, with the sole exception of a partial reduction in the positive percent agreement of the Wake stage. The same result was observed in the comparison between the agreement of the automatic staging with the clinical staging and the agreement of the technicians' staging with the clinical staging. Given these results, the use of this SML-based software may be granted as a supporting tool for sleep clinicians, helping to reduce the burden associated with manual sleep staging of inpatient polysomnography.

In conclusion, as discussed in **Chapter 7,** the studies presented in this doctoral thesis used SML techniques to develop and test algorithms that may support mental health professionals in their clinical practice. Not all studies included in this dissertation present an algorithm that has already reached a definitive, clinically applicable version. However, even at early or intermediate steps, the studies followed specific strategies to provide a preliminary estimate of the generalized performance of the algorithm in order to identify issues in their current versions promptly and better direct the following development steps. The results support the hypothesis that it is possible to develop algorithms that can achieve a

clinically meaningful performance using only input information that would allow an easy clinical translation of these algorithms and avoiding using information that is currently expensive, invasive, or hard to introduce into clinical psychiatric practice. In addition, the studies evidence the crucial role of data, which needs to be of good quality and big-enough quantity to develop SML algorithms. As the use of either purely clinical or experimental data may result in specific issues, data collected in observational, multicentric, non-retrospective studies seem to be the most suitable for SML, and increased availability of open-source datasets with these characteristics may foster the development and test of SML-based clinical tools. Moreover, given that any SML algorithm needs to be transformed into a proper clinical tool before it can be translated into clinical practice, and that the performance that such tool can achieve in practice may be different than what observed for the sole SML algorithm, particular attention should be given also to this development phase, as well as to improve the clinicians' understanding of SML and to remove any resistance from clinicians associated with the utilization of SML-based tools in clinical practice. Only initiatives that promote a coordinated effort between several professional roles and stakeholders, including the end-users of such tools that are clinicians and patients, can finally make SML-based Personalized Medicine clinical tools a widespread reality.

Authors' role in Chapters 2-6

In all studies reported in Chapter 2-6, Massimiliano Grassi contributed to the design of the study, the design and execution of the analyses, the interpretation of results, and the drafting of the manuscript. In the study reported in chapter 4, the co-first Author (Nadine Rouleaux) jointly contributed to the design of the work and the analyses, the interpretation of results and the drafting of the paper. In the study reported in chapter 5, the second Author (Judith Rickelt) jointly contributed to the data preparation and the drafting of the paper. The other Authors contributed either by providing the data used in the analyses (Chapter 2, 3, and 5) or by supervising the studies, from their design to revisiting the manuscripts.

APPENDIX II

CONTRIBUTIONS, IMPACT, AND PROPOSITIONS

Contributions

The contributions of this PhD thesis are:

- the initial development of SML algorithms for the prediction for the 3-year prediction of conversion from Mild Cognitive Impairment to Alzheimer's Disease (Chapter 2-4), and the prediction of 2-year Obsessive-Compulsive Disorder remission (Chapter 5)
- the development of the abovementioned algorithms by using only limited cost and clinically accessible predictive information (Chapter 2-5)
- the introduction of different validation and testing protocols that, even in early development phases, allow providing more accurate estimates of the expected algorithm performance when applied in clinical practice and in multiple clinical contexts, with the aim to steering the development of SML algorithm towards a clinical applicability since the initial phases (Chapter 3-5)
- the extensive clinical validation of a SML algorithm that can automatically perform sleep staging of polysomnography (Chapter 6)

Impact Paragraph

In recent years, there has been an exponential growth of scientific literature regarding the application of SML in Psychiatry. This interest has been motivated by the promise of using SML to develop new clinical tools that could help to perform personalized predictions and recommendations, ultimately improving the results achievable in the psychiatric clinical practice. Starting from the evidence of a substantial lack of such tools in Psychiatry, the studies presented in this dissertation aimed to contribute to further directing the application of SML towards the original promise. In particular, they demonstrate that it is possible to develop SML algorithms that reach clinically relevant performances even by employing only input variables that are or may be easily accessible in the clinical routine, and avoiding those that are still too expensive, invasive, or hard to introduce into clinical psychiatric practice. This selection of the input information is crucial to prevent SML algorithms to

remain just promising proofs-of-concept with limited opportunity to become applicable in practice.

Moreover, the studies also contributed to highlighting the importance of providing estimates of the generalized performance of an SML algorithm even at early development phases. This implies an investigation of what is the expected performance of an SML algorithm when applied in totally new cases, as well as in new clinical settings. This is a necessity because no clinical application of a medical device can be made before a thorough investigation of its safety and efficacy. Doing it systematically at every step of the development process allows to early identify any generalizability issue and to promptly act to solve it along the entire development process. The studies in this thesis introduced peculiar performance testing strategies, specifically designed based on the level at which the development of the algorithm was, the nature of the task under study, and the data available. These strategies may also be used in other studies with similar characteristics or inspire innovative testing protocols.

Overall, the results of the studies included in this doctoral dissertation contribute to demonstrating that the use of SML algorithms in psychiatric clinical practice is not just a promise, even though the process to reach a practical application may require several redesigns of the algorithms and significant evidence in support of their efficacy. Psychiatry may substantially benefit from a shift towards a Personalized Medicine approach to improve the prevention and treatment of mental disorders, which still have significant margins of improvement. Thus, the potential progress achieved in the clinical practice may be worth all the efforts required to complete the development and clinical validation of an SML algorithm. The advantage of using SML is that it does not require an explicit understanding of the phenomena under investigation, but rather the availability of enough suitable examples to be used to train the algorithm. The studies presented in this thesis show how the use of SML may enable to perform psychiatric clinical tasks that were only in part possible previously, e.g., an early prediction of conversion to Alzheimer's Disease in high-risk individuals or of remission in subjects suffering from Obsessive-Compulsive Disorder. The study in Chapter 6 also demonstrates how SML may allow to speed up some clinical procedures

in clinical practice, reducing the costs and the associated clinicians' burden.

This doctoral project has been conceived as inherently multidisciplinary. A joint work among multiple parties and professional figures is necessary to develop SML algorithms that aim to become clinically used tools. Machine Learning requires theoretical and technical skills beyond the average expertise of the typical research scientist in Psychiatry. At the same time, machine learning experts need to work closely and continuously with domain experts from both the research and the clinical side to receive directions regarding which tasks may be relevant to address with SML, which available scientific knowledge can be used to better design and improve the algorithms, and which constrains needs to be satisfied to make them effectively applicable in practice. Besides, further experts need to be involved to effectively transform a SML algorithm into a usable clinical tool, e.g., software engineers, user-experience designers, and regulatory specialists. These interdisciplinary collaborations may foster additional exchanges beyond the sole activities regarding SML, ultimately promoting the beginning of new projects and innovative ideas in all the involved disciplines.

Finally, all studies presented in this doctoral dissertation have been performed with the collaboration of different research groups and institutions, and this doctoral work contributed to further strengthening existing partnerships as well as creating new ones[14]. Part of these collaborations revolved around the sharing of privately-owned datasets that have been used for the development of SML algorithms for the first time. Several psychiatric datasets suitable to be employed for this purpose may exist, but they may not be publicly available, and they may have never been used in SML projects before. The studies presented in this dissertation may also contribute to foster a new use of already available datasets that ultimately will ease the beginning of new SML projects in Psychiatry and make a larger number of institutions and researchers in the psychiatric field approach SML for the first time.

14 The main research groups and institutions involved in this doctoral project were: the School for Mental Health and Neuroscience (MHeNs) and the Institute of Data Science (IDS) of Maastricht University, Villa San Benedetto Menni Hospital (Albese con Cassano, CO, Italy), Humanitas University (Rozzano, MI, Italy), Mount Sinai Medical Center and Miami University (Miami, FL, USA), the Netherlands Obsessive Compulsive Disorder Association (NL), and Medibio Limited (Savage, MN, USA).

Propositions

1. The Personalize Medicine paradigm can help to improve the effectiveness of psychiatric treatments and the prevention of Mental Disorders.
2. Psychiatry must begin to better exploit the possibilities opened by recent technological advancements.
3. The introduction of clinical decision support systems in Psychiatry would allow clinicians to improve the decisions they take in clinical practice.
4. The interest in Psychiatry for Machine Learning (i.e., Artificial Intelligence) needs to be ultimately directed to develop clinical tools.
5. Supervised Machine Learning algorithms should be developed to use input information that can be cost-effectively collected in clinical practice.
6. Introducing the principles of Supervised Machine Learning to mental health clinicians and addressing their potential resistance is necessary to achieve the application of Supervised Machine Learning in clinical practice.
7. A coordinated effort between the academia, industry, clinicians, and patients is necessary to develop useful clinical tools based on Supervised Machine Learning algorithms.
8. Supervised Machine Learning algorithms can significantly contribute to making Personalized Medicine a reality in psychiatric clinical practice.
9. We should not judge the decision-making performance of Artificial Intelligence systems more severely than how we judge the performance of human decision-making.
10. The labor market is drastically changing due to a massive introduction of automatization and Artificial Intelligence, with a serious risk of unemployment for several categories of workers.

APPENDIX III

SUPPLEMENTARY MATERIALS OF CHAPTER 5

Supplementary Table 1. Description of the predictor variables

Source	Variable	Included in the algorithm after training	Descriptive statistics in the train set	Descriptive statistics in the test set	Gain feature importance (score)	Gain feature importance (rank)
ADHD rating scale IV (DuPaul, Ervin et al. 1998)	ADHD symptoms in the past	1	Mean: 4.62, SD: 4.48	Mean: 3.92, SD: 4.21	0.0039	160
	Hyperactivity Impulsiveness symptoms in the past	1	Mean: 2.06, SD: 2.38	Mean: 1.73, SD: 2.21	0.0039	166
	Attention deficit symptoms in the past	1	Mean: 2.57, SD: 2.7	Mean: 2.19, SD: 2.41	0.0035	184
	ADHD inattentive type in the past	1	No: 169 (79%), Yes: 43 (20%), Missing values: 1 (0.47%)	No: 217 (87.15%), Yes: 32 (12.85%)	0.0028	209
	ADHD combined type in the past	1	No: 197 (92%), Yes: 15 (7%), Missing values: 1 (0.47%)	No: 235 (94.38%), Yes: 14 (5.62%)	0.0028	211
	ADHD hyperactive type in the past	0	No: 186 (87%), Yes: 26 (12%), Missing values: 1 (0.47%)	No: 229 (91.97%), Yes: 20 (8.03%)	-	-
Beck Anxiety Inventory (Beck, Epstein et al. 1988)	Becks Anxiety Inventory - total scale score	1	Mean: 15.64, SD: 10.57	Mean: 17.6Yes, SD: 12.06	0.0046	104
Beck Depression Inventory (Beck, Ward et al. 1961)	Becks Depression Inventory - total scale score	1	Mean: 13.63, SD: 8.6	Mean: 15.73, SD: 10.46	0.0059	24
Clinical Interview	How many hours a week the respondent is involved in an executive role in a club or organizations (e.g., sport club, music band, organization for patient, social organization, religious organization, political party)?	1	Mean: 0.17, SD: 0.92	Mean: 0.16, SD: 0.96	0.0076	2
	Antidepressant use on doctor's order in the last two weeks	1	No: 92 (43%), Yes: 110 (52%), Missing values: 11 (5.16%)	No: 75 (30.12%), Yes: 165 (66.27%), Missing values: 9 (4%)	0.0076	3
	Do you have a paid job at the moment?	1	No, I have never had a paid job: 5 (2%), No, I had a paid job in the past: 71 (33%), Yes: 126 (59%), Missing values: 11 (5.16%)	No, I have never had a paid job: 9 (3.61%), No, I had a paid job in the past: 116 (46.59%), Yes: 115 (46.18%), Missing values: 9 (4%)	0.0075	4

Clinical Interview	Psychotropic drug use on doctor's order in the last two weeks	1	No: 80 (38%), Yes: 122 (57%), Missing values: 11 (5.16%)	No: 66 (26.51%), Yes: 174 (69.88%), Missing values: 9 (4%)	0.0070	6
	The respondent participates in a sports club	1	Asked but the respondent did not answer: 5 (2%), No: 134 (63%), Yes: 71 (33%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 165 (66.27%), Yes: 78 (31.33%)	0.0068	7
	Psychoanaleptic use on doctor's order in the last two weeks	1	No: 92 (43%), Yes: 110 (52%), Missing values: 11 (5.16%)	No: 72 (28.92%), Yes: 168 (67.47%), Missing values: 9 (4%)	0.0068	8
	Total number of different psychotropic medications used by the respondent	1	Mean: 0.99, SD: 1.08	Mean: 1.25, SD: 1.2	0.0066	9
	How many hours did you work a week recently?	1	Mean: 16.98, SD: 16.75	Mean: 12.93, SD: 16.37	0.0064	10
	How many different antidepressants are used by the respondent?	1	Mean: 0.6, SD: 0.64	Mean: 0.76, SD: 0.58	0.0064	11
	Age at the time of the interview	1	Mean: 39.95, SD: 10.75	Mean: 36.47, SD: 11.02	0.0063	13
	Antipsychotic use on doctor's order the last two weeks	1	No: 183 (86%), Yes: 19 (9%), Missing values: 11 (5.16%)	No: 196 (78.71%), Yes: 44 (17.67%), Missing values: 9 (4%)	0.0061	20
	How many different psychoanaleptic medications are used by the respondent?	1	Mean: 0.62, SD: 0.65	Mean: 0.79, SD: 0.6	0.0060	21
	Do you take classes aimed at a diploma at the moment or in the last year?	1	Asked but the respondent did not answer: 7 (3%), No: 185 (87%), Yes: 18 (8%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), No: 217 (87.15%), Yes: 27 (10.84%)	0.0058	27
	The respondent participates in a political party, organization or club	1	Asked but the respondent did not answer: 5 (2%), No: 194 (91%), Yes: 11 (5%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 230 (92.37%), Yes: 13 (5.22%)	0.0058	28
	Number of children	1	Mean: 0.4Yes, SD: 0.91	Mean: 0.45, SD: 0.93	0.0057	32
	How many different anxiolytic medications are used by the respondent?	1	Mean: 0.13, SD: 0.34	Mean: 0.14, SD: 0.39	0.0057	33

Clinical Interview	*How often do you have contact (phone, email, writing a letter, etc) with your best friend?*	1	Asked but the respondent did not answer: 10 (5%), No friend: 54 (25%), Less than a few times a year: 2 (1%), A few times a year: 18 (8%), A few times a month: 61 (29%), A few times a week: 54 (25%), Daily: 10 (5%), We live in the same house: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 14 (5.62%), No friend: 57 (22.89%), Less than a few times a year: 5 (2.01%), A few times a year: 19 (7.63%), A few times a month: 63 (25.3%), A few times a week: 67 (26.91%), Daily: 15 (6.02%), We live in the same house: 4 (1.61%), Missing values: 5 (2%)	0.0057	34
	Participant currently taking serotonergic antidepressiant according to clinical guidelines for OCD	1	Missing: 15 (7%), None: 97 (46%), Yes, dosage not reported: (1%), Yes, subtherapeutic dosage according to guidelines: 38 (18%), Yes, adequate OCD dosage according to guidelines: 60 (28%)	Missing: 14 (5.62%), None: 79 (31.73%), Yes, subtherapeutic dosage according to guidelines: 75 (30.12%), Yes, adequate OCD dosage according to guidelines: 81 (32.53%)	0.0055	44
	Psycholeptic use on doctor's order in the last two weeks	1	No: 158 (74%), Yes: 44 (21%), Missing values: 11 (5.16%)	No: 175 (70.28%), Yes: 65 (26.1%), Missing values: 9 (4%)	0.0054	49
	How often does the respondent visit a sport match?	1	Asked but the respondent did not answer: 9 (4%), Practically never: 164 (77%), A few times a year: 24 (11%), Every month: 4 (2%), A few times a month: 5 (2%), Every week: 2 (1%), A few times a week: 2 (1%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), Practically never: 179 (71.89%), A few times a year: 30 (12.05%), Every month: 9 (3.61%), A few times a month: 12 (4.82%), Every week: 12 (4.82%), A few times a week: 1 (0.4%)	0.0053	52
	How many minutes a week do you volunteer for a club or organizations (e.g., sport club, music band, organization for patient, social organization, religious organization, political party)?	1	Mean: 0.52, SD: 3.02	Mean: 1.37, SD: 5.75	0.0053	54

Clinical Interview	Total household income of the respondent (excluding holiday allowance, refunding of traveling or payment of expense)	1	Asked, no answer: 26 (12%), 0-750 euro: 14 (7%), 0750-1000 euro: 18 (8%), 1000-1250 euro: 18 (8%), 1250-1500 euro: 21 (10%), 1500-2000 euro: 25 (12%), 2000-2500 euro: 11 (5%), 2500-3000 euro: 24 (11%), 3000-4000 euro: 28 (13%), more than 4000 euro: 14 (7%), Missing values: 14 (6.57%)	Asked, no answer: 45 (18.07%), 0-750 euro: 7 (2.81%), 0750-1000 euro: 21 (8.43%), 1000-1250 euro: 17 (6.83%), 1250-1500 euro: 19 (7.63%), 1500-2000 euro: 26 (10.44%), 2000-2500 euro: 21 (8.43%), 2500-3000 euro: 30 (12.05%), 3000-4000 euro: 34 (13.65%), more than 4000 euro: 18 (7.23%), Missing values: 11 (4%)	0.0052	59
	Of the friends you have, how many are people you work with?	1	Mean: 0.34, SD: 1.23	Mean: 0.36, SD: 1.26	0.0052	61
	Does the respondent have an executive role in a club or organizations (e.g., sport club, music band, organization for patient, social organization, religious organization, political party)?	1	Asked but the respondent did not answer: 6 (3%), No: 189 (89%), Yes: 15 (7%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 7 (2.81%), No: 219 (87.95%), Yes: 23 (9.24%)	0.0051	62
	How often do you see\visit your best friend?	1	Asked but the respondent did not answer: 6 (3%), No friend: 54 (25%), Less than a few times a year: 1 (0%), A few times a year: 52 (24%), A few times a month: 56 (26%), A few times a week: 35 (16%), Daily: 5 (2%), We live in the same house: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 3 (1.2%), No friend: 57 (22.89%), Less than a few times a year: 5 (2.01%), A few times a year: 46 (18.47%), A few times a month: 76 (30.52%), A few times a week: 45 (18.07%), Daily: 8 (3.21%), We live in the same house: 4 (1.61%), Missing values: 5 (2%)	0.0051	66
	How many hours a day do you spend with hobbies, doing "odd" jobs or other creative activities around the house?	1	Mean: 0.86, SD: 1.29	Mean: 0.85, SD: 1.32	0.0050	69
	How among your friends are neighbors or live in the neighborhood?	1	Mean: 1.26, SD: 1.9	Mean: 1.2, SD: 2.28	0.0050	70

Clinical Interview	Employment status of the respondent	1	Incapacitated for work: 42 (20%), Paid work, 12 hours a week or more: 98 (46%), Paid work, but less than 12 hours a week: 4 (2%), Retired: 5 (2%), Housewife / house husband: 6 (3%), Student: 12 (6%), Unemployed: 7 (3%), Working as a volunteer: 5 (2%), Independent worker: 9 (4%), Independent worker: 14 (7%), Missing values: 11 (5.16%)	Incapacitated for work: 67 (26.91%), Paid work, 12 hours a week or more: 89 (35.74%), Paid work, but less than 12 hours a week: 7 (2.81%), Retired: 4 (1.61%), Housewife / house husband: 10 (4.02%), Student: 20 (8.03%), Unemployed: 7 (2.81%), Working as a volunteer: 9 (3.61%), Independent worker: 5 (2.01%), Sickness Benefits Act: 22 (8.84%), Missing values: 9 (4%)	0.0050	72
	How many minutes a day do you spend with hobbies, doing odd jobs or other creative activities around the house?	1	Mean: 7.44, SD: 12.55	Mean: 7.86, SD: 12.55	0.0050	73
	Total number of different benzodiazepines used by the respondent	1	Mean: 0.19, SD: 0.44	Mean: 0.18, SD: 0.48	0.0050	75
	How many friends do you have or how many friends do you think you have?	1	Mean: 5.5, SD: 5.26	Mean: 5.28, SD: 4.77	0.0050	77
	How many times a week you read the newspaper?	1	Asked but the respondent did not answer: 9 (4%), Never: 46 (22%), Less than once a week: 21 (10%), 1-2 times a week: 33 (15%), 3-4 times a week: 31 (15%), Daily: 70 (33%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), Never: 86 (34.54%), Less than once a week: 26 (10.44%), 1-2 times a week: 46 (18.47%), 3-4 times a week: 20 (8.03%), Daily: 66 (26.51%)	0.0049	88
	Do you have a partner at the moment?	1	questionnaire not present: 14 (7%), no: 70 (33%), yes: 129 (61%)	questionnaire not present: 11 (4.42%), no: 86 (34.54%), yes: 152 (61.04%)	0.0048	91

Clinical Interview	Living arrangements of the respondent	1	Other: 9 (4%), Alone: 74 (35%), Partner with childeren: 49 (23%), Partner without childeren: 45 (21%), Single with childeren: 8 (4%), With parents: 10 (5%), Missing values: 18 (8.45%)	Other: 9 (3.61%), Alone: 68 (27.31%), Partner with childeren: 68 (27.31%), Partner without childeren: 63 (25.3%), Single with childeren: 5 (2.01%), With parents: 25 (10.04%), Missing values: 11 (4%)	0.0048	93
	Do you use a computer?	1	Asked but the respondent did not answer: 7 (3%), No: 28 (13%), Yes: 175 (82%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), No: 28 (11.24%), Yes: 216 (86.75%)	0.0048	95
	Type of residence of the respondent	1	Other: 2 (1%), In parents house: 8 (4%), Lodgings: 7 (3%), Own house (rented or owner-occupied): 121 (57%), Not pertinent: 58 (27%), Missing values: 17 (7.98%)	Other: 8 (3.21%), In parents house: 18 (7.23%), Lodgings: 7 (2.81%), Own house (rented or owner-occupied): 140 (56.22%), Not pertinent: 64 (25.7%), Missing values: 12 (5%)	0.0047	96
	How many hours a week do you use a computer?	1	Mean: 6.65, SD: 9.93	Mean: 7.42, SD: 10.0	0.0047	98
	If you participate in clubs or organizations, how often do you attend activities or meetings of these clubs or organizations?	1	Asked but the respondent did not answer: 6 (3%), Never: 104 (49%), Practically never: 4 (2%), A few times a year: 21 (10%), Every month: 10 (5%), A few times a month: 11 (5%), Every week: 28 (13%), A few times a week: 25 (12%), Every day: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 7 (2.81%), Never: 105 (42.17%), Practically never: 6 (2.41%), A few times a year: 23 (9.24%), Every month: 12 (4.82%), A few times a month: 17 (6.83%), Every week: 41 (16.47%), A few times a week: 37 (14.86%), Every day: 1 (0.4%)	0.0046	101
	Number of benzodiazepines from different groups used by the respondent	1	Mean: 0.18, SD: 0.42	Mean: 0.17, SD: 0.42	0.0046	102
	Body Mass Index (BMI) of the respondent	1	Mean: 24.7, SD: 4.8	Mean: 25.95, SD: 5.67	0.0045	112

Clinical Interview	How often does the respondent go to a cafe, restaurant, etc.?	1	Asked but the respondent did not answer: 9 (4%), Practically never: 26 (12%), A few times a year: 48 (23%), Every month: 45 (21%), A few times a month: 33 (15%), Every week: 33 (15%), A few times a week: 15 (7%), Every day: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), Practically never: 38 (15.26%), A few times a year: 64 (25.7%), Every month: 51 (20.48%), A few times a month: 48 (19.28%), Every week: 37 (14.86%), A few times a week: 5 (2.01%), Every day: 1 (0.4%)	0.0045	114
	The respondent participates in a Trade Union, a employers' organization, or a professional organization	1	Asked but the respondent did not answer: 5 (2%), No: 186 (87%), Yes: 19 (9%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 224 (89.96%), Yes: 19 (7.63%)	0.0044	116
	How many different psycholeptic medications are used by the respondent?	1	Mean: 0.28, SD: 0.6	Mean: 0.4, SD: 0.77	0.0044	118
	How many different benzodiazepines are used by the respondent?	1	Mean: 0.13, SD: 0.34	Mean: 0.14, SD: 0.39	0.0044	121
	How often do you watch the news or the newsreel on TV?	1	Asked but the respondent did not answer: 9 (4%), Never: 10 (5%), 2 Less than once a week: 14 (7%), 3 1-2 times a week: 20 (9%), 4 3-4 times a week: 41 (19%), 5 Daily: 116 (54%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), Never: 21 (8.43%), 2 Less than once a week: 15 (6.02%), 3 1-2 times a week: 25 (10.04%), 4 3-4 times a week: 55 (22.09%), 5 Daily: 127 (51%)	0.0044	122
	Is your best friend a man or a woman?	1	Asked but the respondent did not answer: 6 (3%), No best friend: 54 (25%), Man: 55 (26%), Woman: 95 (45%), Missing values: 3 (1.41%)	No best friend: 57 (22.89%), Man: 76 (30.52%), Woman: 111 (44.58%), Missing values: 5 (2%)	0.0043	123
	How many persons in your household have a regular income? Don't include children with only Saturday/holidays jobs	1	Mean: 1.39, SD: 0.65	Mean: 1.6Yes, SD: 0.78	0.0043	124
	If you participate in clubs or organizations, do you volunteer for these clubs or organizations?	1	Asked but no answer: 5 (2%), No: 168 (79%), Yes: 37 (17%), values: 3 (1.41%)	Asked but the respondent did not answer: 7 (2.81%), No: 178 (71.49%), Yes: 64 (25.7%)	0.0043	127

Clinical Interview	How many different SSRIs are used by the respondent?	1	Mean: 0.37, SD: 0.49	Mean: 0.48, SD: 0.52	0.0042	131
	How often does the respondent go to the forest, dunes, zoo, etc.?	1	Asked but the respondent did not answer: 7 (3%), Practically never: 34 (16%), A few times a year: 59 (28%), Every month: 32 (15%), A few times a month: 35 (16%), Every week: 26 (12%), A few times a week: 10 (5%), Every day: 7 (3%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), Practically never: 46 (18.47%), A few times a year: 98 (39.36%), Every month: 37 (14.86%), A few times a month: 24 (9.64%), Every week: 20 (8.03%), A few times a week: 10 (4.02%), Every day: 9 (3.61%)	0.0042	134
	Sex of the respondent	1	Male: 95 (45%), Female: 118 (55%)	Male: 119 (47.79%), Female: 130 (52.21%)	0.0042	136
	Did the respondent experience one or more negative events in the past year?	1	No: 54 (25%), Yes: 159 (75%)	No: 54 (21.69%), Yes: 191 (76.71%), Missing values: 4 (2%)	0.0042	139
	How often does the respondent go shopping for fun?	1	Asked but the respondent did not answer: 7 (3%), Practically never: 43 (20%), A few times a year: 38 (18%), Every month: 47 (22%), A few times a month: 40 (19%), Every week: 24 (11%), A few times a week: 10 (5%), Every day: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 7 (2.81%), Practically never: 60 (24.1%), A few times a year: 49 (19.68%), Every month: 44 (17.67%), A few times a month: 45 (18.07%), Every week: 33 (13.25%), A few times a week: 10 (4.02%), Every day: 1 (0.4%)	0.0041	149
	Do you take classes at the moment or in last year?	1	Asked but the respondent did not answer: 7 (3%), No: 138 (65%), Yes: 65 (31%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), No: 161 (64.66%), Yes: 83 (33.33%)	0.0041	151
	How often does the respondent visit a cultural organization, like a movie theater, museum, concert, etc.?	1	Practically never: 44 (21%), A few times a year: 93 (44%), Every month: 34 (16%), A few times a month: 21 (10%), Every week: 7 (3%), A few times a week: 4 (2%), Missing values: 3 (1.41%)	Practically never: 71 (28.51%), A few times a year: 133 (53.41%), Every month: 26 (10.44%), A few times a month: 12 (4.82%), Every week: 2 (0.8%)	0.0040	153

	Number of positive recent events in the past year	1	Mean: 2.2, SD: 1.63	Mean: 2.16, SD: 1.62	0.0040	154
	How many different antipsychotic medications are used by the respondent?	1	Mean: 0.1Yes, SD: 0.37	Mean: 0.22, SD: 0.51	0.0040	158
	Did the respondent experience one or more lingering conflicts or problems in the past year?	1	No: 81 (38%), Yes: 132 (62%)	No: 115 (46.18%), Yes: 129 (51.81%), Missing values: 5 (2%)	0.0039	159
	Do you have financial troubles?	1	No: 174 (82%), Yes: 24 (11%), Missing values: 15 (7.04%)	No: 215 (86.35%), Yes: 23 (9.24%), Missing values: 11 (4%)	0.0039	161
	Number of negative recent events in the past year	1	Mean: 1.63, SD: 1.5	Mean: 1.64, SD: 1.43	0.0039	164
	If you participate in clubs or organizations, how many minutes a week are you involved in an executive role in these clubs or organizations?	1	Mean: 0.19, SD: 2.16	Mean: 1.08, SD: 5.6	0.0039	167
	Other antidepressant (different than SSRIs and non-selective monoamine reuptake inhibitors) use on doctor's order in the last two weeks	1	No: 174 (82%), Yes: 28 (13%), Missing values: 11 (5.16%)	No: 220 (88.35%), Yes: 20 (8.03%), Missing values: 9 (4%)	0.0038	168
Clinical Interview	How often does the respondent visit a social-cultural organization?	1	Asked but the respondent did not answer: 9 (4%), Practically never: 164 (77%), A few times a year: 13 (6%), Every month: 10 (5%), A few times a month: 4 (2%), Every week: 9 (4%), A few times a week: 1 (0%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), Practically never: 204 (81.93%), A few times a year: 17 (6.83%), Every month: 7 (2.81%), A few times a month: 6 (2.41%), Every week: 6 (2.41%), A few times a week: 4 (1.61%)	0.0038	169
	How often does the respondent do outdoor activities, like swimming, walking, fishing, etc.?	1	Practically never: 33 (15%), A few times a year: 13 (6%), Every month: 11 (5%), A few times a month: 15 (7%), Every week: 51 (24%), A few times a week: 55 (26%), Every day: 25 (12%), Missing values: 3 (1.41%)	Practically never: 56 (22.49%), A few times a year: 27 (10.84%), Every month: 10 (4.02%), A few times a month: 25 (10.04%), Every week: 55 (22.09%), A few times a week: 45 (18.07%), Every day: 25 (10.04%)	0.0038	171
	Benzodiazepines (all types) use on doctor's order the last two weeks	1	No: 168 (79%), Yes: 34 (16%), Missing values: 11 (5.16%)	No: 205 (82.33%), Yes: 35 (14.06%), Missing values: 9 (4%)	0.0038	174

Clinical Interview	Number of lingering conflicts or problems in the past year	1	Mean: 0.85, SD: 0.79	Mean: 0.74, SD: 0.81	0.0037	177
	The respondent participates in other kind of clubs and organizations	1	Asked but the respondent did not answer: 5 (2%), No: 179 (84%), Yes: 26 (12%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 220 (88.35%), Yes: 23 (9.24%)	0.0035	183
	The respondent participates in a choral society, music band, theatre company	1	Asked but the respondent did not answer: 5 (2%), No: 182 (85%), Yes: 23 (11%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 223 (89.56%), Yes: 20 (8.03%)	0.0035	185
	Did the respondent experience one or more positive recent events in the past year?	1	No: 34 (16%), Yes: 179 (84%)	No: 40 (16.06%), Yes: 205 (82.33%), Missing values: 4 (2%)	0.0034	188
	The respondent participates in an organization for patients	1	Asked but the respondent did not answer: 5 (2%), No: 183 (86%), Yes: 22 (10%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 217 (87.15%), Yes: 26 (10.44%)	0.0034	189
	SSRI use on doctor's order the last two weeks	1	No: 128 (60%), Yes: 74 (35%), Missing values: 11 (5.16%)	No: 127 (51%), Yes: 113 (45.38%), Missing values: 9 (4%)	0.0034	190
	Participant currently taking psychotropic pharmacotherapy according to guidelines	1	missing: 18 (8%), no: 135 (63%), yes: 60 (28%)	missing: 13 (5.22%), no: 154 (61.85%), yes: 81 (32.53%), Missing values: 1 (0%)	0.0034	192
	If you participate in clubs or organizations, how many hours a week do you volunteer for these clubs and organizations?	1	Mean: 0.6Yes, SD: 2.43	Mean: 0.88, SD: 2.62	0.0033	193
	Anxiolytic use on doctor's order the last two weeks	1	No: 175 (82%), Yes: 27 (13%), Missing values: 11 (5.16%)	No: 210 (84.34%), Yes: 30 (12.05%), Missing values: 9 (4%)	0.0033	194
	If you participate in clubs or organizations, do you participate in activities or meetings of these clubs or organizations?	1	Asked but the respondent did not answer: 5 (2%), No: 104 (49%), Yes: 101 (47%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 105 (42.17%), Yes: 138 (55.42%)	0.0033	195
	How many different non-selective monoamine reuptake inhibitors are used by the respondent?	1	Mean: 0.09, SD: 0.32	Mean: 0.19, SD: 0.43	0.0032	199

Clinical Interview	The respondent participates in an action committee or organization with social goals	1	Asked but the respondent did not answer: 5 (2%), No: 186 (87%), Yes: 19 (9%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 232 (93.17%), Yes: 11 (4.42%)	0.0030	203
	How many minutes a week do you use a computer?	1	Mean: 3.08, SD: 9.37	Mean: 5.59, SD: 12.24	0.0030	204
	Do you take other kind of classes at the moment or in the last year?	1	Asked but the respondent did not answer: 7 (3%), No: 168 (79%), Yes: 35 (16%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 5 (2.01%), No: 206 (82.73%), Yes: 38 (15.26%)	0.0030	205
	The respondent participates in a church or organization with religious or ideological goal	1	Asked but the respondent did not answer: 5 (2%), No: 186 (87%), Yes: 19 (9%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 177 (71.08%), Yes: 66 (26.51%)	0.0029	206
	The respondent participates in a hobby or social club	1	Asked but the respondent did not answer: 5 (2%), No: 187 (88%), Yes: 18 (8%), Missing values: 3 (1.41%)	Asked but the respondent did not answer: 6 (2.41%), No: 222 (89.16%), Yes: 21 (8.43%)	0.0029	207
	Benzodiazepine (anxiolytic type) use on doctor's order in the last two weeks	1	No: 175 (82%), Yes: 27 (13%), Missing values: 11 (5.16%)	No: 210 (84.34%), Yes: 30 (12.05%), Missing values: 9 (4%)	0.0029	208
	Non-selective monoamine reuptake inhibitor use on doctor's order in the last two weeks	1	No: 185 (87%), Yes: 17 (8%), Missing values: 11 (5.16%)	No: 197 (79.12%), Yes: 43 (17.27%), Missing values: 9 (4%)	0.0025	216
	How many different hypnotics sedatives are used by the respondent?	1	Mean: 0.04, SD: 0.22	Mean: 0.03, SD: 0.18	0.0010	217
	How many different antiepileptic medications are used by the respondent?	0	Mean: 0.03, SD: 0.23	Mean: 0.02, SD: 0.14	-	-
	How many different hypnotic benzodiazepines are used by the respondent?	0	Mean: 0.03, SD: 0.2	Mean: 0.03, SD: 0.18	-	-
	How many different nervous system drugs are used by the respondent??	0	Mean: 0.0Yes, SD: 0.12	Mean: 0.0Yes, SD: 0.14	-	-
	How many different addictive disorders drug are used by the respondent??	0	Mean: 0.0Yes, SD: 0.12	Mean: 0.0Yes, SD: 0.14	-	-
EuroQol (EuroQol 1990)	EQ-5D score	1	Mean: 0.7Yes, SD: 0.26	Mean: 0.68, SD: 0.27	0.0057	31

Interpretation of Intrusion Inventory (Group 2001)	Interpretation of Intrusions Inventory: Responsibility subscale score	1	Mean: 448.26, SD: 277.36	Mean: 488.28, SD: 281.01	0.0047	99
	Interpretation of Intrusions Inventory: Importance of Thoughts subscale score	1	Mean: 359.0Yes, SD: 228.58	Mean: 371.73, SD: 246.24	0.0046	100
	Interpretation of Intrusions Inventory: Control subscale score	1	Mean: 519.52, SD: 246.79	Mean: 554.96, SD: 260.81	0.0041	147
Level of Expressed Emotion (Cole and Kazarian 1988)	Percieved lack of emotional support scale	1	Mean: 31.02, SD: 9.52	Mean: 31.69, SD: 12.11	0.0050	76
	Percieved irritation scale	1	Mean: 12.9, SD: 4.64	Mean: 13.27, SD: 5.01	0.0049	79
	Percieved intrusiveness scale	1	Mean: 11.77, SD: 5.31	Mean: 12.6Yes, SD: 5.79	0.0049	89
	Percieved criticism scale	1	Mean: 8.45, SD: 2.79	Mean: 8.78, SD: 3.15	0.0047	97
Life Chart (Eaton, Anthony et al. 1997)	Chronical Course of OCD in the last 2 years	1	Too many omitted answers from the respondent: 3 (1%), No: 90 (42%), Yes: 120 (56%)	Too many omitted answers from the respondent: 3 (1.2%), No: 97 (38.96%), Yes: 146 (58.63%), Missing values: 3 (1%)	0.0070	5
	Late onset OCD (20 years or older)?	1	Asked but the respondent did not answer: 11 (5%), No: 126 (59%), Yes: 59 (28%), Missing values: 17 (7.98%)	Asked but the respondent did not answer: 15 (6.02%), No: 144 (57.83%), Yes: 87 (34.94%), Missing values: 3 (1%)	0.0057	30
Loneliness Scale (De Jong-Gierveld and Kamphuls 1985)	Emotional loneliness score	1	Mean: 2.66, SD: 2.18	Mean: 2.69, SD: 2.22	0.0056	38
	Total score	1	Mean: 5.18, SD: 3.6	Mean: 5.42, SD: 3.61	0.0050	74
	Social loneliness score	1	Mean: 2.53, SD: 1.86	Mean: 2.73, SD: 1.82	0.0049	85
Montgomery-Asberg Depression Rating Scale (Montgomery and Asberg 1979)	MADRS total score	1	Mean: 11.92, SD: 8.09	Mean: 12.04, SD: 9.3	0.0062	14
Padua Inventory (Sanavio 1988)	Padua Inventory: precision subscale score	1	Mean: 6.45, SD: 5.9	Mean: 6.75, SD: 6.2	0.0062	15
	Padua Inventory: rumination subscale score	1	Mean: 20.85, SD: 9.5	Mean: 22.84, SD: 8.89	0.0051	63
	Padua Inventory: checking subscale score	1	Mean: 13.07, SD: 7.91	Mean: 13.94, SD: 7.16	0.0051	65
	Padua Inventory: washing subscale score	1	Mean: 11.87, SD: 11.92	Mean: 10.73, SD: 10.16	0.0051	67
	Padua Inventory: impulses subscale score	1	Mean: 6.07, SD: 5.98	Mean: 5.62, SD: 6.52	0.0050	68
Structured Clinical Interview for DSM-IV-R (First and Gibbon 2004)	Diagnosis of Somatoform disorders - lifetime	1	No: 203 (95%), Yes: 10 (5%)	No: 236 (94.78%), Yes: 13 (5.22%)	0.0064	12

Structured Clinical Interview for DSM-IV-R (First and Gibbon 2004)	Diagnosis of Major depressive disorder - lifetime	1	No: 83 (39%), Yes: 130 (61%)	No: 96 (38.55%), Yes: 153 (61.45%)	0.0052	58
	Diagnosis of Specific Phobia - current	1	No: 189 (89%), Yes: 24 (11%)	No: 240 (96.39%), Yes: 9 (3.61%)	0.0049	84
	Any current diagnosis of Anxiety disorder besides OCD diagnosis	1	No: 140 (66%), Yes: 73 (34%)	No: 172 (69.08%), Yes: 77 (30.92%)	0.0046	103
	Any lifetime diagnosis of anxiety disorder besides OCD diagnosis	1	No: 112 (53%), Yes: 101 (47%)	No: 138 (55.42%), Yes: 111 (44.58%)	0.0046	109
	Diagnosis of Social phobia - lifetime	1	No: 161 (76%), Yes: 52 (24%)	No: 179 (71.89%), Yes: 70 (28.11%)	0.0044	117
	Number of current diagnosis	1	Mean: 1.84, SD: 1.11	Mean: 1.77, SD: 1.02	0.0042	129
	Diagnosis of Dysthymic disorder - lifetime	1	No: 197 (92%), Yes: 16 (8%)	No: 234 (93.98%), Yes: 15 (6.02%)	0.0042	137
	Diagnosis of Panic disorder with agoraphobia - lifetime	1	No: 177 (83%), Yes: 36 (17%)	No: 226 (90.76%), Yes: 23 (9.24%)	0.0042	142
	Number of lifetime diagnosis	1	Mean: 2.74, SD: 1.46	Mean: 2.63, SD: 1.36	0.0041	150
	Diagnosis of Substance related disorders dependence - lifetime	1	No: 194 (91%), Yes: 19 (9%)	No: 222 (89.16%), Yes: 27 (10.84%)	0.0040	157
	Diagnosis of Social phobia - current	1	No: 182 (85%), Yes: 31 (15%)	No: 204 (81.93%), Yes: 45 (18.07%)	0.0038	172
	Diagnosis of Specific Phobia - lifetime	1	No: 179 (84%), Yes: 34 (16%)	No: 230 (92.37%), Yes: 19 (7.63%)	0.0036	180
	Diagnosis of Eating disorders - lifetime	1	No: 186 (87%), Yes: 27 (13%)	No: 223 (89.56%), Yes: 26 (10.44%)	0.0035	187
	Diagnosis of Major depressive disorder - current	1	No: 175 (82%), Yes: 38 (18%)	No: 210 (84.34%), Yes: 39 (15.66%)	0.0032	197
	Diagnosis of Generalized anxiety disorder - lifetime	1	No: 187 (88%), Yes: 26 (12%)	No: 228 (91.57%), Yes: 21 (8.43%)	0.0032	198
	Diagnosis of Panic disorder without agoraphobia - lifetime	1	No: 194 (91%), Yes: 19 (9%)	No: 222 (89.16%), Yes: 27 (10.84%)	0.0032	200
	Diagnosis of Dysthymic disorder - current	1	No: 202 (95%), Yes: 11 (5%)	No: 239 (95.98%), Yes: 10 (4.02%)	0.0028	210
	Diagnosis of Panic disorder with agoraphobia - current	1	No: 200 (94%), Yes: 13 (6%)	No: 238 (95.58%), Yes: 11 (4.42%)	0.0027	214
	Diagnosis of Generalized anxiety disorder - current	0	No: 195 (92%), Yes: 18 (8%)	No: 231 (92.77%), Yes: 18 (7.23%)	-	-

Self-reported general attachment style (Griffin and Bartholomew 1994)	*Which attachment style most appropriately describes you?*	1	Asked but the respondent did not answer: 4 (2%), Dismissing: 21 (10%), Fearful: 65 (31%), Preoccupied: 52 (24%), Secure: 66 (31%), Missing values: 5 (2.35%)	Asked but the respondent did not answer: 9 (3.61%), Dismissing: 21 (8.43%), Fearful: 104 (41.77%), Preoccupied: 48 (19.28%), Secure: 57 (22.89%), Missing values: 10 (4%)	0.0052	60
	Attachment Style Fearful score: *I am wary to get engaged in close relationships because I am afraid to get hurt*	1	Mean: 4.05, SD: 2.01	Mean: 4.38, SD: 1.88	0.0059	25
	Attachment Style Preoccupied score: *I have the impression that usually I like others better than they like me*	1	Mean: 3.74, SD: 1.77	Mean: 3.9, SD: 1.68	0.0055	42
	Attachment Style Dismissing score: *I prefer that others are independent of me and I am independent of them*	1	Mean: 3.24, SD: 1.86	Mean: 2.96, SD: 1.85	0.0049	86
	Attachment Style Secure score: *I feel at ease in intimate relationships*	1	Mean: 4.08, SD: 1.83	Mean: 3.7Yes, SD: 1.78	0.0042	138
Social Support Inventory (Timmerman, Emanuels-Zuurveen et al. 2000)	Informative Support subscale score	1	Mean: 12.84, SD: 2.22	Mean: 12.55, SD: 2.46	0.0046	110
	Instrumental Support subscale score	1	Mean: 13.25, SD: 2.19	Mean: 13.16, SD: 2.51	0.0042	141
	Emotional Support subscale score	1	Mean: 12.6, SD: 2.65	Mean: 12.2Yes, SD: 2.75	0.0041	144
	Social Companionship subscale score	1	Mean: 12.Yes, SD: 2.56	Mean: 12.15, SD: 2.68	0.0041	145
Structured Trauma Interview (Draijer and Langeland 1999)	Mother was (sometimes) dysfuctioning or unavailable?	1	Too many omitted answers from the respondent: 5 (2%), No: 90 (42%), Yes: 112 (53%), Missing values: 6 (2.82%)	No: 101 (40.56%), Yes: 148 (59.44%)	0.0060	22
Structured Trauma Interview (Draijer and Langeland 1999)	Mother and/or Father was (sometimes) dysfuctioning or unavailable	1	Too many omitted answers from the respondent: 5 (2%), No: 57 (27%), Yes: 145 (68%), Missing values: 6 (2.82%)	No: 72 (28.92%), Yes: 177 (71.08%)	0.0059	23
	Physical abuse (domestic) after age 16	1	Asked but the respondent did not answer: 30 (14%), No: 156 (73%), Yes: 21 (10%), Missing values: 6 (2.82%)	Asked but the respondent did not answer: 28 (11.24%), No: 186 (74.7%), Yes: 35 (14.06%)	0.0057	29

	Abuse before or after age 16	1	Asked but the respondent did not answer: 30 (14%), No: 164 (77%), Yes: 13 (6%), Missing values: 6 (2.82%)	Asked but the respondent did not answer: 29 (11.65%), No: 206 (82.73%), Yes: 14 (5.62%)	0.0057	36
	Father was (sometimes) dysfuctioning or unavailable?	1	Too many omitted answers from the respondent: 9 (4%), No: 101 (47%), Yes: 97 (46%), Missing values: 6 (2.82%)	No: 136 (54.62%), Yes: 113 (45.38%)	0.0054	46
	Total score dysfunctioning or unavailability mother	1	Mean: 1.14, SD: 1.31	Mean: 1.16, SD: 1.31	0.0050	71
	Childhood witnessing of interparental violence	1	Asked but the respondent did not answer: 5 (2%), No: 167 (78%), Yes: 35 (16%), Missing values: 6 (2.82%)	Asked but the respondent did not answer: 5 (2.01%), No: 212 (85.14%), Yes: 32 (12.85%)	0.0049	80
Structured Trauma Interview (Draijer and Langeland 1999)	Number of questions unanswered	1	Mean: 0.25, SD: 1.02	Mean: 0.02, SD: 0.14	0.0049	87
	Total score dysfunctioning or unavailability father	1	Mean: 0.94, SD: 1.21	Mean: 0.78, SD: 1.12	0.0048	90
	Number of different kind of childhood trauma exposures before age 16 (0-6: mother and father disfunctioning is counted separately)	1	Mean: 1.5Yes, SD: 1.22	Mean: 1.4Yes, SD: 1.17	0.0046	105
	Number of different kind of childhood trauma exposures before age 16 (0-5: mother and father disfunctioning is counted together)	1	Mean: 1.2, SD: 0.96	Mean: 1.07, SD: 0.88	0.0044	119
	Sexual abuse after age 16	1	Asked but the respondent did not answer: 13 (6%), No: 162 (76%), Yes: 32 (15%), Missing values: 6 (2.82%)	Asked but the respondent did not answer: 5 (2.01%), No: 210 (84.34%), Yes: 34 (13.65%)	0.0044	120
	Physical abuse but no sexual abuse before age 16	1	No: 188 (88%), Yes: 19 (9%), Missing values: 6 (2.82%)	No: 231 (92.77%), Yes: 18 (7.23%)	0.0043	125
	Sexual and/or physical abuse before age 16	1	No: 180 (85%), Yes: 27 (13%), Missing values: 6 (2.82%)	No: 218 (87.55%), Yes: 31 (12.45%)	0.0040	155
	Physical (parental) abuse before age 16	1	No: 186 (87%), Yes: 21 (10%), Missing values: 6 (2.82%)	No: 228 (91.57%), Yes: 21 (8.43%)	0.0026	215
Systolic and diastolic blood pressure assessment	Diastolic pressure - arm - lying - measurement 1	1	Mean: 79.12, SD: 10.75	Mean: 79.89, SD: 12.78	0.0054	50
	Systolic pressure - arm - lying - measurement 2	1	Mean: 131.17, SD: 18.93	Mean: 131.12, SD: 17.25	0.0052	57

Systolic and diastolic blood pressure assessment	Diastolic pressure - arm - standing - measurement 2	1	Mean: 83.74, SD: 10.52	Mean: 84.4Yes, SD: 12.65	0.0050	78
	Systolic pressure - arm - standing - measurement 1	1	Mean: 128.88, SD: 17.61	Mean: 129.9Yes, SD: 16.57	0.0049	82
	Diastolic pressure - arm - lying - measurement 2	1	Mean: 79.37, SD: 10.94	Mean: 81.04, SD: 13.28	0.0048	92
	Systolic pressure - arm - standing - measurement 2	1	Mean: 130.69, SD: 18.62	Mean: 131.83, SD: 18.38	0.0046	107
	Systolic pressure - arm - lying - measurement 1	1	Mean: 131.9Yes, SD: 17.5	Mean: 131.26, SD: 17.2	0.0046	108
	Diastolic pressure - arm - standing - measurement 1	1	Mean: 82.16, SD: 10.59	Mean: 83.17, SD: 11.41	0.0041	152
Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	Respondent doing houshold work	1	Did not do it, because of health problems: 11 (5%), Did not do it, for reasons other than health problems: 3 (1%), Done, hindered by health problems: 118 (55%), Done, not hindered by health problems: 81 (38%)	Did not do it, because of health problems: 6 (2.41%), Did not do it, for reasons other than health problems: 4 (1.61%), Done, hindered by health problems: 156 (62.65%), Done, not hindered by health problems: 80 (32.13%), Missing values: 3 (1%)	0.0062	16
	Hours of work missed/lost because of hindrance by health problems	1	Mean: 19.25, SD: 92.91	Mean: 10.15, SD: 53.65	0.0062	18
	With how many medical specialists did you have contact in the last 6 months?	1	Mean: 0.63, SD: 0.91	Mean: 0.54, SD: 0.84	0.0061	19
	Have you been admitted to a health care institution in the last 6 months?	1	No: 189 (89%), Yes: 24 (11%)	No: 169 (67.87%), Yes: 77 (30.92%), Missing values: 3 (1%)	0.0058	26
	Days per week the respondent is employed	1	Mean: 2.33, SD: 2.12	Mean: 1.86, SD: 2.23	0.0057	35
	I was at work, but due to health problems I had to postpone work for the past 6 months	1	Asked but the respondent did not answer: 1 (0%), Not pertinent: 129 (61%), Rarely: 43 (20%), Occasionally: 10 (5%), Sometimes: 22 (10%), Often: 6 (3%), Nearly all the time: 2 (1%)	Asked but the respondent did not answer: 2 (0.8%), Not pertinent: 177 (71.08%), Rarely: 43 (17.27%), Occasionally: 5 (2.01%), Sometimes: 12 (4.82%), Often: 5 (2.01%), Nearly all the time: 2 (0.8%), Missing values: 3 (1%)	0.0056	37

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	Respondent doing “odd” jobs	1	Did not do it, because of health problems: 36 (17%), Did not do it, for reasons other than health problems: 38 (18%), Done, hindered by health problems: 76 (36%), Done, not hindered by health problems: 63 (30%)	Did not do it, because of health problems: 29 (11.65%), Did not do it, for reasons other than health problems: 56 (22.49%), Done, hindered by health problems: 96 (38.55%), Done, not hindered by health problems: 65 (26.1%), Missing values: 3 (1%)	0.0056	39
	Number of hours volunteers took over domestic work in past 6 months	1	Mean: 4.13, SD: 37.8	Mean: 3.63, SD: 24.14	0.0056	40
	If the respondent had contact with a psychiatrist, psychologist or psychotherapist in a policlinic of a general hospital in the last six months, in what type of hospital did the respondent have such contact?	1	General Hospital: 4 (2%), No contact with a psychiatrist, psychologist or psychotherapist in a policlinic of a general hospital: 119 (56%), Other type of hospital: 17 (8%), Psychiatric hospital: 63 (30%), University hospital: 9 (4%)	General Hospital: 10 (4.02%), No contact with a psychiatrist, psychologist or psychotherapist in a policlinic of a general hospital: 185 (74.3%), Other type of hospital: 4 (1.61%), Psychiatric hospital: 21 (8.43%), University hospital: 20 (8.03%), Missing values: 3 (1%)	0.0055	41
	Hours per week the respondent is employed	1	Mean: 17.67, SD: 16.14	Mean: 13.7Yes, SD: 16.44	0.0055	43
	If the respondent does not work, what is the best description of current status?	1	(early) Retirement: 4 (2%), Housekeeping: 6 (3%), No work because of health related problems: 56 (26%), Other reasons: 8 (4%), Not pertinent: 134 (63%), Student: 5 (2%)	(early) Retirement: 5 (2.01%), Housekeeping: 15 (6.02%), No work because of health related problems: 88 (35.34%), Other reasons: 7 (2.81%), Not pertinent: 118 (47.39%), Student: 13 (5.22%), Missing values: 3 (1%)	0.0055	45
	How many contacts with a social worker in the last six months?	1	Mean: 0.53, SD: 3.82	Mean: 1.14, SD: 6.91	0.0054	48
	Did you have any contact with a social worker in the last six months?	1	No: 201 (94%), Yes: 12 (6%)	No: 226 (90.76%), Yes: 19 (7.63%), Missing values: 4 (2%)	0.0054	51

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	*Do you have a paid job at the moment?*	1	No: 79 (37%), Yes: 134 (63%)	No: 128 (51.41%), Yes: 118 (47.39%), Missing values: 3 (1%)	0.0053	53
	Did you have any contact with a psychiatrist, psychologist or psychotherapist in a policlinic of a general hospital without admission to the hospital in the last six months?	1	No: 119 (56%), Yes: 94 (44%)	No: 185 (74.3%), Yes: 58 (23.29%), Missing values: 6 (2%)	0.0053	55
	How many contacts did you have with psychiatrist, psychologist or psychotherapist in policlinic of a general hospital without admission to the hospital in the last six months?	1	Mean: 3.43, SD: 7.41	Mean: 2.66, SD: 7.29	0.0052	56
	Did you have any contact with a RIAGG or GGZ institute in the last six months?	1	No: 113 (53%), Yes: 100 (47%)	No: 88 (35.34%), Yes: 156 (62.65%), Missing values: 5 (2%)	0.0051	64
	Has the respondent being absent from work due to health problems in last 6 months?	1	No: 142 (67%), Yes: 70 (33%), Missing values: 1 (0.47%)	No: 169 (67.87%), Yes: 76 (30.52%), Missing values: 4 (2%)	0.0049	81
	I was at work, but due to health problems I had problems with concentration in the past 6 months	1	Not pertinent: 129 (61%), Rarely: 18 (8%), Occasionally: 9 (4%), Sometimes: 31 (15%), Often: 16 (8%), Nearly all the time: 9 (4%), Missing values: 1 (0.47%)	Not pertinent: 177 (71.08%), Rarely: 26 (10.44%), Occasionally: 10 (4.02%), Sometimes: 17 (6.83%), Often: 11 (4.42%), Nearly all the time: 4 (1.61%), Missing values: 4 (2%)	0.0049	83
	How many contacts did you have with a RIAGG or GGZ institute in the last six months?	1	Mean: 12.62, SD: 26.94	Mean: 13.8, SD: 29.02	0.0048	94
	If the respondent has children, did the respondent do things for or with the children living at home?	1	Asked but the respondent did not answer/not relevant: 104 (49%), Did not do it, because of health problems: 1 (0%), Did not do it, for reasons other than health problems: 47 (22%), Done, hindered by health problems: 32 (15%), Done, not hindered by health problems: 29 (14%)	Asked but the respondent did not answer/not relevant: 74 (29.72%), Did not do it, because of health problems: 5 (2.01%), Did not do it, for reasons other than health problems: 86 (34.54%), Done, hindered by health problems: 43 (17.27%), Done, not hindered by health problems: 38 (15.26%), Missing values: 3 (1%)	0.0046	106

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	*I was at work, but due to health problems I had to work at a slower pace over the past 6 months*	1	Not pertinent: 129 (61%), Rarely: 28 (13%), Occasionally: 9 (4%), Sometimes: 20 (9%), Often: 20 (9%), Nearly all the time: 6 (3%), Missing values: 1 (0.47%)	Not pertinent: 177 (71.08%), Rarely: 23 (9.24%), Occasionally: 9 (3.61%), Sometimes: 22 (8.84%), Often: 7 (2.81%), Nearly all the time: 7 (2.81%), Missing values: 4 (2%)	0.0045	111
	Volunteers took over domestic work of the respondent in past 6 months?	1	No: 190 (89%), Yes: 22 (10%), Missing values: 1 (0.47%)	No: 227 (91.16%), Yes: 18 (7.23%), Missing values: 4 (2%)	0.0045	113
	Did you have any contact with a medical specialist in a policlinic of a general hospital without admission to the hospital in the last six months?	1	No: 121 (57%), Yes: 91 (43%), Missing values: 1 (0.47%)	No: 152 (61.04%), Yes: 92 (36.95%), Missing values: 5 (2%)	0.0044	115
	I was at work, but due to health problems I had to isolate myself for the past 6 months	1	Not pertinent: 129 (61%), Rarely: 51 (24%), Occasionally: 8 (4%), Sometimes: 15 (7%), Often: 8 (4%), Nearly all the time: 1 (0%), Missing values: 1 (0.47%)	Not pertinent: 177 (71.08%), Rarely: 43 (17.27%), Occasionally: 5 (2.01%), Sometimes: 16 (6.43%), Often: 2 (0.8%), Nearly all the time: 2 (0.8%), Missing values: 4 (2%)	0.0043	126
	How many contacts did you have with a physiotherapist in the last six months?	1	Mean: 6.72, SD: 22.91	Mean: 2.69, SD: 7.89	0.0043	128
	In what type of health care institution have you been admitted?	1	General Hospital: 14 (7%), No admission to a health care institution : 189 (89%), Other type of hospital: 1 (0%), Psychiatric hospital: 4 (2%), University hospital: 5 (2%)	General Hospital: 17 (6.83%), No admission to a health care institution : 169 (67.87%), Other type of hospital: 26 (10.44%), Psychiatric hospital: 29 (11.65%), University hospital: 2 (0.8%), Missing values: 6 (2%)	0.0042	130
	Physical or psychological cause of absence/illness/disability?	1	Not at all: 129 (61%), A little: 59 (28%), A lot: 24 (11%), Missing values: 1 (0.47%)	Not at all: 177 (71.08%), A little: 46 (18.47%), A lot: 22 (8.84%), Missing values: 4 (2%)	0.0042	132
	How many contacts did you have with your physician in the last six months? (add all visits to doctor, telephonic consultations, and visits of the physician at the respondent's home)	1	Mean: 1.95, SD: 2.05	Mean: 1.94, SD: 2.24	0.0042	133

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	*Did you participate in a self-help group in the last six months? (e.g., Aa group, talk-group patient association)?*	1	No: 196 (92%), Yes: 16 (8%), Missing values: 1 (0.47%)	No: 231 (92.77%), Yes: 14 (5.62%), Missing values: 4 (2%)	0.0042	135
	Did you have any contact with alternative caretakers in the last six months (like a homoeopath, acupuncturist, healer, manual therapist, haptonomist chiropractor, iriscopist?)	1	No: 173 (81%), Yes: 40 (19%)	No: 219 (87.95%), Yes: 26 (10.44%), Missing values: 4 (2%)	0.0042	140
	How many days have you been admitted to a health care institution in the last 6 months?	1	Mean: 4.36, SD: 31.62	Mean: 21.3Yes, SD: 59.36	0.0042	143
	Did family members took over domestic work in past 6 months?	1	No: 152 (71%), Yes: 60 (28%), Missing values: 1 (0.47%)	No: 179 (71.89%), Yes: 67 (26.91%), Missing values: 3 (1%)	0.0041	146
	Respondent going to buy Groceries	1	Did not do it, because of health problems: 7 (3%), Did not do it, for reasons other than health problems: 4 (2%), Done, hindered by health problems: 88 (41%), Done, not hindered by health problems: 114 (54%)	Did not do it, because of health problems: 9 (3.61%), Did not do it, for reasons other than health problems: 4 (1.61%), Done, hindered by health problems: 127 (51%), Done, not hindered by health problems: 106 (42.57%), Missing values: 3 (1%)	0.0041	148
	Number of hours family members took over domestic work in past 6 months	1	Mean: 23.92, SD: 89.67	Mean: 19.94, SD: 61.76	0.0040	156
	Days in the last 6 months the respondent was hindered at work by health problems	1	Mean: 19.7Yes, SD: 38.12	Mean: 20.18, SD: 43.59	0.0039	162
	Did you have any contact with a physiotherapist in the last six months?	1	No: 146 (69%), Yes: 66 (31%), Missing values: 1 (0.47%)	No: 196 (78.71%), Yes: 49 (19.68%), Missing values: 4 (2%)	0.0039	163
	I was at work, but due to health problems I had to have others take over work for the past 6 months	1	Not pertinent: 129 (61%), Rarely: 54 (25%), Occasionally: 10 (5%), Sometimes: 12 (6%), Often: 7 (3%), Missing values: 1 (0.47%)	Not pertinent: 177 (71.08%), Rarely: 47 (18.88%), Occasionally: 7 (2.81%), Sometimes: 9 (3.61%), Often: 5 (2.01%), Missing values: 4 (2%)	0.0039	165
	How many whole days did you have day-time or part-time treatment for mental problems in the last 6 months?	1	Mean: 4.3, SD: 22.92	Mean: 10.84, SD: 30.31	0.0038	170

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	*Did you have any contact with a company doctor in the last six months?*	1	No: 164 (77%), Yes: 49 (23%)	Asked but the respondent did not answer: 9 (3.61%), No: 179 (71.89%), Yes: 58 (23.29%), Missing values: 3 (1%)	0.0038	173
	I was at work, but due to health issues, I had more trouble making decisions in the past 6 months	1	Asked but the respondent did not answer: 2 (1%), Not pertinent: 129 (61%), Rarely: 43 (20%), Occasionally: 9 (4%), Sometimes: 20 (9%), Often: 9 (4%), Nearly all the time: 1 (0%)	Asked but the respondent did not answer: 1 (0.4%), Not pertinent: 177 (71.08%), Rarely: 39 (15.66%), Occasionally: 9 (3.61%), Sometimes: 10 (4.02%), Often: 7 (2.81%), Nearly all the time: 3 (1.2%), Missing values: 3 (1%)	0.0038	175
	How many contacts did you have with a company doctor in the last six months?	1	Mean: 0.73, SD: 1.75	Mean: 0.88, SD: 2.11	0.0038	176
	Others took over domestic work that the respondent normally does in past 6 months?	1	No: 130 (61%), Yes: 83 (39%)	No: 159 (63.86%), Yes: 87 (34.94%), Missing values: 3 (1%)	0.0037	178
	I was at work, but due to health problems I had other problems	1	Not pertinent: 129 (61%), Rarely: 60 (28%), Occasionally: 2 (1%), Sometimes: 9 (4%), Often: 7 (3%), Nearly all the time: 5 (2%)	Not pertinent: 178 (71.49%), Rarely: 51 (20.48%), Sometimes: 4 (1.61%), Often: 8 (3.21%), Nearly all the time: 3 (1.2%), Missing values: 3 (1%)	0.0037	179
	With how many alternative caretakers did you have contact in the last 6 months?	1	Mean: 0.2Yes, SD: 0.45	Mean: 0.12, SD: 0.36	0.0036	181
	How many contacts did you have with an independent psychiatrist, psychologist or psychotherapist in the last six months?	1	Mean: 2.65, SD: 10.02	Mean: 2.06, SD: 11.2	0.0036	182
	Did you have any contact with an independent psychiatrist, psychologist or psychotherapist in the last six months?	1	No: 159 (75%), Yes: 53 (25%), Missing values: 1 (0.47%)	No: 208 (83.53%), Yes: 36 (14.46%), Missing values: 5 (2%)	0.0035	186
	How many different self-help groups?	1	No: 197 (92%), Yes: 16 (8%)	No: 235 (94.38%), Yes: 14 (5.62%)	0.0034	191
	Hours per week the respondent was employed in the past	1	Mean: 6.46, SD: 13.51	Mean: 9.78, SD: 16.19	0.0033	196
	Did you have contact with your physician in the last six months?	1	No: 49 (23%), Yes: 164 (77%)	No: 72 (28.92%), Yes: 174 (69.88%), Missing values: 3 (1%)	0.0032	201

Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (Roijen, Straten et al. 2002)	*In what type of institution did you have daytime- or part-time treatment for mental problems?*	1	General Hospital: 1 (0%), No daytime- or parttime treatment for mental problems: 190 (89%), Other type of hospital: 15 (7%), Psychiatric hospital: 7 (3%)	Asked but the respondent did not answer: 6 (2.41%), General Hospital: 2 (0.8%), No daytime- or parttime treatment for mental problems: 183 (73.49%), Other type of hospital: 22 (8.84%), Psychiatric hospital: 33 (13.25%), Missing values: 3 (1%)	0.0031	202
	Did you have day-time or part-time treatment for mental problems?	1	No: 190 (89%), Yes: 23 (11%)	No: 183 (73.49%), Yes: 63 (25.3%), Missing values: 3 (1%)	0.0028	212
	How many contacts for a homecare did you have *in the last 6 months?*	1	Mean: 7.65, SD: 49.34	Mean: 7.73, SD: 51.37	0.0028	213
	Number of hours homecare took over domestic work in past 6 months	0	Mean: 1.89, SD: 11.98	Mean: 1.07, SD: 10.37	-	-
	Number of hours paid help took over domestic work in past 6 months	0	Mean: 3.77, SD: 38.18	Mean: 2.92, SD: 15.35	-	-
	How many contacts did you have with the center for alcohol and drugs in the last six months?	0	Mean: 0.24, SD: 1.78	Mean: 0.55, SD: 4.91	-	-
	Did you use homecare in the last six months?	0	No: 202 (95%), Yes: 11 (5%)	No: 229 (91.97%), Yes: 16 (6.43%), Missing values: 4 (2%)	-	-
Y-BOCS (Goodman, Price et al. 1989, Goodman, Price et al. 1989)	Total severity score	1	Mean: 19.96, SD: 6.94	Mean: 19.9Yes, SD: 7.35	0.0077	1
	Total severity score - compulsions	1	Mean: 10.26, SD: 4.28	Mean: 10.33, SD: 4.28	0.0062	17
	Total severity score - obsessions	1	Mean: 9.95, SD: 3.83	Mean: 10.06, SD: 4.09	0.0054	47

The Italic font in the variable description indicates questions asked to the respondent. For some of these questions, minor adaptations have been made to the text reported in this table in order to improve their understandability.

References

Beck, A. T., N. Epstein, G. Brown and R. A. Steer (1988). "An inventory for measuring clinical anxiety: psychometric properties." J Consult Clin Psychol **56**(6): 893-897.

Beck, A. T., C. H. Ward, M. Mendelson, J. Mock and J. Erbaugh (1961). "An inventory for measuring depression." Arch Gen Psychiatry **4**: 561-571.

Cole, J. D. and S. S. Kazarian (1988). "The level of expressed emotion scale: a new measure of expressed emotion." Journal of Clinical psychology **44**(3): 392-397.

De Jong-Gierveld, J. and F. Kamphuls (1985). "The development of a Rasch-type loneliness scale." Applied psychological measurement **9**(3): 289-299.

Draijer, N. and W. Langeland (1999). "Childhood trauma and perceived parental dysfunction in the etiology of dissociative symptoms in psychiatric inpatients." American Journal of Psychiatry **156**(3): 379-385.

DuPaul, G. J., R. A. Ervin, C. L. Hook and K. E. McGoey (1998). "Peer tutoring for children with attention deficit hyperactivity disorder: effects on classroom behavior and academic performance." J Appl Behav Anal **31**(4): 579-592.

Eaton, W. W., J. C. Anthony, J. Gallo, G. Cai, A. Tien, A. Romanoski, C. Lyketsos and L. S. Chen (1997). "Natural history of Diagnostic Interview Schedule/DSM-IV major depression. The Baltimore Epidemiologic Catchment Area follow-up." Arch Gen Psychiatry **54**(11): 993-999.

EuroQol, G. (1990). "EuroQol--a new facility for the measurement of health-related quality of life." Health Policy **16**(3): 199-208.

First, M. B. and M. Gibbon (2004). The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID-II). Comprehensive handbook of

psychological assessment, Vol. 2: Personality assessment. Hoboken, NJ, US, John Wiley & Sons Inc**:** 134-143.

Goodman, W. K., L. H. Price, S. A. Rasmussen, C. Mazure, P. Delgado, G. R. Heninger and D. S. Charney (1989). "The Yale-Brown Obsessive Compulsive Scale. II. Validity." Arch Gen Psychiatry **46**(11): 1012-1016.

Goodman, W. K., L. H. Price, S. A. Rasmussen, C. Mazure, R. L. Fleischmann, C. L. Hill, G. R. Heninger and D. S. Charney (1989). "The Yale-Brown Obsessive Compulsive Scale. I. Development, use, and reliability." Arch Gen Psychiatry **46**(11): 1006-1011.

Griffin, D. W. and K. Bartholomew (1994). "Models of the self and other: Fundamental dimensions underlying measures of adult attachment." Journal of personality and social psychology **67**(3): 430.

Group, O. C. C. W. (2001). "Development and initial validation of the obsessive beliefs questionnaire and the interpretation of intrusions inventory." Behaviour Research and Therapy **39**(8): 987-1006.

Montgomery, S. A. and M. Asberg (1979). "A New Depression Scale Designed to be Sensitive to Change." Br J Psychiatry **134**: 382-389.

Roijen, L., A. Straten, B. Tiemens and M. Donker (2002). "Manual Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (TIC-P) (in Dutch)." **2**.

Sanavio, E. (1988). "Obsessions and compulsions: the Padua Inventory." Behav Res Ther **26**(2): 169-177.

Timmerman, I., E. Emanuels-Zuurveen and P. Emmelkamp (2000). "The Social Support Inventory (SSI): A Brief Scale to Assess Perceived Adequacy of Social Support." Clinical Psychology & Psychotherapy **7**: 401-410.

APPENDIX IV

ACKNOWLEDGMENTS

Science is never a lonely journey. This is part of its beauty (and fun). All achievements are team achievements and there are so many who made all this work possible. I want to thank all of you, but in particular some special people I have been sharing this journey with.

Prof. dr. Koen Schruers, for making it possible in the first place, for his supervision and patience, for his constant help, bright advice, and support in all these years since the Masters in Affective Neurosciences. Duizendmaal dank!

Prof. dr. Giampaolo Perna, for having been my mentor and leader since the last year of my bachelor's in Psychology (it was long time ago, at the beginning of 2005). His seemingly simple question on the first day I joined his lab, "Beyond psychology, what do you also like?" and my answer to it, "Well, I also like physics and enjoyed the statistics course" were laterally the turning point that directed me to where I am nowadays.

Prof. dr. Michel Dumontier for having seen potential in me without knowing me at all in the beginning. After years with great guidance from the clinical and research science fields, while being basically me alone with the literature and the peer-to-peer support of the community from the technical side, Prof. Dumontier was finally able to fill this gap.

Great Professors always have great teams behind, and I cannot do without starting to thank the whole Prof. dr. G. Perna's team, my real science buddies for more than 15 years. Like every band, members may come and go but one has always been there: thank you dr. Daniela Caldirola, in particular for your tireless tutoring along all these years.

I have also been sharing my journey with two other excellent teams at Maastricht University: Prof. dr. K. Schruers' team at the School for Mental Health and Neuroscience and Department of Psychiatry and Neuropsychology, and Prof. dr. M. Dumontier's Institute of Data Science. Both were essential in helping me get out when my point of view was "stuck in a local minimum", grow personally and professionally, support me in thousands of ways, and add fun to my Maastricht visits.

I also want to thank the whole team at Medibio Limited, and Archie Defillo in particular. Eventually, we want all our efforts to change the life of people suffering from Mental Disorders for the better. My collaboration with Medibio has been bringing me firsthand to the last mile (actually, it would be better to say several miles) that separate science from making patients finally profit from science achievements. A really invaluable experience that has been contributing to shaping my scientific work.

For the same reason, I want to thank all patients I have worked with in my clinical practice. The burden of their suffering, sometimes shamefully kept inside, sometimes misunderstood and criticized by other people, fulfills all this of meaning.

Last, but definitely not the least, I want to thank my family, and especially and deeply my wife Laura. For all the motivation and support I received to get this work completed in all these years, despite the huge amount of time it required to be stolen from our family.

APPENDIX V

CURRICULUM VITAE

Date and Place of Birth: 11/10/1983, Treviglio (BG), Italy.

Education

May 2016
Machine Learning Summer School. Cadiz, Spain.

January 2009 – May 2014
Post-graduate specialization in Cognitive Psychotherapy. Centro Studi Cognitivi (Como, Italy). Grade: 70/70.

July 2008 - July 2009 / July 2011 – July 2012
Master of Science (M.Sc.), Affective neuroscience. Maastricht University (The Netherlands), Grade: 8,2/10.

September 2006
Visiting Student. Stanford University / VA Hospital, Palo Alto, CA, USA.

October 2005– September 2007
Master of Science (M.Sc.), Clinical Psychology.
Vita-Salute San Raffaele University (Milan, Italy). Grades: 110/110 with honors.
Thesis project: Development of two supervised machine learning algorithms (Artificial Neural Networks) as tools to support clinical decision making in Psychiatry.

October 2002 – September 2005
Bachelor of Science (B.Sc.), Psychological Sciences.
Vita-Salute San Raffaele University (Milan, Italy). Grades: 110/110 with honors.
Thesis on the application of Artificial Neural Networks in Psychiatry.

September 1997 – July 2002
Scientific High School.
Liceo Statale Galielo Galilei, Caravggio (BG), Italy.

Working Experience

March 2019 – nowadays
Head of Artificial Intelligence (since February 2020), formerly **Principal Data Scientist** at Medibio Ltd, USA/AU (https://medibio.com.au)

September 2009 – nowadays
Senior Data Scientist, Research Scientist at Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.
As part of this job (September 2016 – nowadays)**:**

- Collaborations as **data scientist (machine learning)**: Institute of Data Science, Maastricht University, Netherlands; Department of Psychiatry and Neuropsychology, Faculty of Health, Medicine and Life Sciences, University of Maastricht, Netherlands; Center of Aging, Department of Psychiatry and Behavioral Sciences, Miller School of Medicine, University of Miami, USA
- **Collaborator (statistical analyses)** in the project "Are Anxiety Disorders associated with accelerated cognitive decline and molecular mechanisms of dementia? A multi-centric Italian study in middle-aged and older patients and controls", founded by the Cariplo Foundation, Italy.

May 2019 – April 2020
Consultant Researcher at Humanitas University, Rozzano, Milan, Italy.

September 2014 - September 2015
Consultant Statistician in the R.O.I. (Italian Osteopath Registry) Schools Multi-Centric Project.

March 2004 – June 2009
Research Assistant at Department of Clinical Neuroscience, San Raffaele Turro Hospital and
Vita Salute San Raffaele University, Milan, Italy.

July 2019 – nowadays
Clinical Psychologist at Center for Personalized Medicine of Panic and Anxiety Disorders, Humanitas San Pio X Hospital, Milan, Italy

May 2010 – nowadays
Clinical Psychologist at CEDANS (Villa San Benedetto Hospital outpatient facilities), Albese con Cassano (CO), Italy, and Milan, Italy.

Teaching Experience

February 2019
Lecturer in Machine Learning for Clinical Psychology at the AffectTech Horizon 2020 Marie Skłodowska-Curie Innovative Training Networks Program, Cattolica University, Milan, Italy.

October 2018 – nowadays
Adjunct Professor in Communication Skills at the Degree Course in Medicine and Surgery,
International Medical School, Faculty of Medicine, Humanitas University, Rozzano, Milan, Italy.

September 2018
Lecturer in Machine Learning at the postgraduate Master in Bioinformatics and Functional Genomics, University of Milan, Milan, Italy.

October 2012 – nowadays
Lecturer in Statistics at ISO School of Osteopathic Medicine, Milan, Italy. B.Sc. and M.Sc. degrees in Osteopathy granted by Bucks New University, High Wycombe, UK. Previously, B.SC. degrees granted by University of Wales, Cardiff, UK.

October 2016 – December 2019
Mentor for the course "Machine Learning: Regression" at Coursera (http://www.coursera.org).

February 2018
Lecturer in the CME accredited course "From Data Analysis to Planning in Healthcare" at the Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.

December 2014 – December 2016
Lecturer in Methodology of Research at the Music Therapy School, Lucio Campani Music Conservatory, Mantova, Italy

October 2013 – May 2014
Lecturer in Psychology at ISO School of Osteopathic Medicine, Milan, Italy. B.Sc. and M.Sc. degrees in Osteopathy granted by Bucks New University, High Wycombe, UK. Previously, B.SC. degrees granted by University of Wales, Cardiff, UK.

November 2011– June 2012
Lecturer in the CME accredited course "Scientific Method and Literature Update" at the Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.

September 2010 – November 2010
Lecturer in the CME accredited course "Introduction to Statistics and Scientific Method" at the Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.

January 2007 – December 2007
Teaching Assistant in the online course "Expert in non-pharmacological psychiatric interventions", modules "Introduction to Statistics" and "Introduction to Functional Psychopathology", funded by the European Social Fund (project n° 238568), Centro Formazione Professionale Padre Monti, Saronno (VR), Italy.

Editorial Experience

January 2018 – December 2018
Associate Editor of the Journal of Alzheimer's Disease.

2014 – nowadays
Peer-reviewer for various scientific journals (e.g., Health and Quality of Life Outcomes, Journal of Health Psychology, Current Bioinformatics, Personalized Medicine in Psychiatry)

Grants and Fellowship

February 2018 – April 2018
Research Fellowship in "Alzheimer's Disease" granted by the International Foundation for the Support of Research in Psychiatry (http://www.fondazioneforipsi.org). Research activity carried out at the Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.

September 2009 – August 2012
Research Fellowship in "Functional Psychopathology of Psychiatric Disorders" granted by the International Foundation for the Support of Research in Psychiatry (http://www.fondazioneforipsi.org). Research activity carried out at the Department of Clinical Neuroscience, Villa San Benedetto Hospital, Albese con Cassano (CO), Italy.

July 2011
Spinoza Grant, scholarship for the International master in Affective Neuroscience. European Accreditation Committee in CNS (EACIC), academic year 2011-2012.

November 2007 – October 2008
Research Fellowship in "Systems in Support of Clinical Decision Making" at Department of Clinical Neuroscience, San Raffaele Turro Hospital and Vita Salute San Raffaele University, Milan, Italy.

APPENDIX VI

PUBLICATIONS

Publications in Scientific Journals

Caldirola, D., Daccò, S., Cuniberti, F., Grassi, M., Alciati, A., Torti, T., Perna, G. “First-onset major depression during the COVID-19 pandemic: A predictive machine learning model”. J Affect Disord **310**:75-86.

Caldirola, D., F. Cuniberti, S. Daccò, M. Grassi, T. Torti and G. Perna. “Predicting new-onset psychiatric disorders throughout the COVID-19 pandemic: A machine learning approach”. J Neuropsychiatry Clin Neurosci in press.

Grassi, M., J.Rickelt, D. Caldirola, M. Eikelenboom, P. van Oppen, M. Dumontier, G. Perna and K. Schruers (2022). “Prediction of illness remission in patients with Obsessive-Compulsive Disorder with supervised machine learning”. J Affect Disord **296**:117–125.

Fernández-Álvarez, J., M. Grassi, D. Colombo, D. Colombo, C. Botella, P. Cipresso, G. Perna and G. Riva (2021). "The Efficacy of Bio- and Neurofeedback for Depression: A Meta-analysis." Psychol Med **15**:1-16.

Caldirola, D., S. Daccò, F. Cuniberti, M. Grassi, S. Lorusso, G. Diaferia and G. Perna (2021). “Elevated C-reactive protein levels across diagnoses: The first comparison among inpatients with major depressive disorder, bipolar disorder, or obsessive-compulsive disorder”. J Psychosomatic Res **150**:110604.

Perna, G., F. Cuniberti, S. Daccò, M. Grassi and D. Caldirola (2020). "‘Precision’or ‘personalized’psychiatry: different terms–same content?" Fortschritte der Neurologie· Psychiatrie.

Perna, G., A. Alciati, S. Daccò, M. Grassi and D. Caldirola (2021). "Personalized psychiatry and depression: the role of sociodemographic and clinical variables." Psychiatry Investigation **17**(3): 193.

Grassi, M., N. Rouleaux, D. Caldirola, D. Loewenstein, K. Schruers, G. Perna and M. Dumontier (2019). "A Novel Ensemble-Based Machine Learning Algorithm to Predict the Conversion From Mild Cognitive Impairment to Alzheimer's Disease Using Socio-Demographic Characteristics, Clinical Information, and Neuropsychological Measures." Front Neurol **10**: 756.

Perna, G., E. Sangiorgio, M. Grassi and D. Caldirola (2018). "Commentary: cognitive behavioral therapy vs. eye movement desensitization and reprocessing for treating panic disorder: a randomized controlled trial." Frontiers Psychol **9**: 1061.

Grassi, M., D. A. Loewenstein, D. Caldirola, K. Schruers, R. Duara and G. Perna (2018). "A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion: further evidence of its accuracy via a transfer learning approach." Int. Psychogeriatr. **14**: 1-9.

Alciati, A., F. Atzeni, M. Grassi, D. Caldirola, P. Sarzi-Puttini, J. Angst and G. Perna (2018). "Features of mood associated with high body weight in females with fibromyalgia." Comprehensive Psychiatry **80**: 57-64.

Caldirola, D., E. Sangiorgio, A. Riva, M. Grassi, A. Alciati, C. Scialò and G. Perna (2017). "Does gender influence cognitive function in non-psychotic depression?" Personalized Medicine in Psychiatry **4**: 25-31.

Caldirola, D., M. Grassi, A. Alciati, A. Riva, E. Sangiorgio, S. Daccò and G. Perna (2017). "Personalized medicine in panic disorder: Where are we now? A meta-regression analysis." Personalized Medicine in Psychiatry **1**: 26-38.

Alciati, A., D. Caldirola, M. Grassi, D. Foschi and G. Perna (2017). "Mediation effect of recent loss events on weight gain in obese people who experienced childhood parental death or separation." Journal of Health Psychology **22**(1): 101-110.

Alciati, A., F. Atzeni, M. Grassi, D. Caldirola, A. Riva, P. Sarzi-Puttini and G. Perna (2017). "Childhood adversities in patients with fibromyalgia: are they related to comorbid lifetime major depression?" Clinical and experimental rheumatology **35**(3): 112-118.

Caldirola, D., M. Grassi, A. Riva, S. Daccò, D. De Berardis, B. Dal Santo and G. Perna (2014). "Self-reported quality of life and clinician-rated functioning in mood and anxiety disorders: relationships and neuropsychological correlates." Comprehensive psychiatry **55**(4): 979-988.

Grassi, M., D. Caldirola, N. V. Di Chiaro, A. Riva, S. Daccò, M. Pompili and G. Perna (2014). "Are respiratory abnormalities specific for panic disorder? A meta-analysis." Neuropsychobiology **70**(1): 52-60.

Riva, A., P. Cavedini, G. Guerriero, D. Prestia, M. Grassi and G. Perna (2013). "2169–Does nicotine have a pro-cognitive effect in obsessive-compulsive disorder?" European Psychiatry **28**: 1.

Grassi, M., D. Caldirola, G. Vanni, G. Guerriero, M. Piccinni, A. Valchera and G. Perna (2013). "Baseline respiratory parameters in panic disorder: a meta-analysis." Journal of affective disorders **146**(2): 158-173.

Caldirola, D., S. Daccò, M. Grassi, A. Citterio, R. Menotti, P. Cavedini, P. Girardi and G. Perna (2013). "Effects of cigarette smoking on neuropsychological performance in mood disorders: a comparison between smoking and nonsmoking inpatients." The Journal of clinical psychiatry **74**(2): 130-136.

Perna, G., D. Di Pasquale, M. Grassi, G. Vanni, L. Bellodi and D. Caldirola (2012). "Temperament, character and anxiety sensitivity in panic disorder: a high-risk study." Psychopathology **45**(5): 300-304.

Caldirola, D., R. Teggi, S. Bondi, F. L. Lopes, M. Grassi, M. Bussi and G. Perna (2011). "Is there a hypersensitive visual alarm system in panic disorder?" Psychiatry research **187**(3): 387-391.

Favaron, E., D. Caldirola, M. Grassi, S. Biffi, E. Galimberti, L. Bellodi and G. Perna (2006). "Response to paroxetine and polymorphism of the Catechol-O-ethyl transferase in panic disorder." Journal of Psychopharmacology **20**(4): S109-S109.

Book Chapters

Chiera, M., M. Grassi (2018). “Statistica di base” (“Elements of Statistics”), in Cerritelli F., D. Lanaro. “Elementi di ricerca in osteopatia e terapia manuale”. Edises, Neaples, Italy.