Temporal prediction of multiple sclerosis evolution from patient-centered outcomes

Samuele Fiorini SAMUELE.FIORINI@DIBRIS.UNIGE.IT
Alessandro Verri ALESSANDRO.VERRI@UNIGE.IT
Annalisa Barla ANNALISA.BARLA@UNIGE.IT
Department of Informatics, Bioengineering, Robotics and System Engineering (DIBRIS)
Università degli Studi di Genova
Via Dodecaneso 35, 16146, Genova, Italy

Andrea Tacchino ANDREA.TACCHINO@AISM.IT
Giampaolo Brichetto GIAMPAOLO.BRICHETTO@AISM.IT
Scientific Research Area
Italian Multiple Sclerosis Foundation
Via Operai 40, 16149, Genova, Italy

Abstract

Multiple Sclerosis is a degenerative condition of the central nervous system that affects nearly 2.5 million of individuals in terms of their physical, cognitive, psychological and social capabilities. Despite the high variability of its clinical presentation, *relapsing* and *progressive* multiple sclerosis are considered the two main disease types, with the former possibly evolving into the latter. Recently, the attention of the medical community toward the use of patient-centered outcomes in multiple sclerosis has significantly increased. Such patient-friendly measures are devoted to the assessment of the impact of the disease on several domains of the patient life. In this work, we investigate on use of patient-centered outcomes to predict the evolution of the disease and to assess its impact on patients' lives. To this aim, we build a novel temporal model based on gradient boosting classification and multiple-output elastic-net regression. The model provides clinically interpretable results along with accurate predictions of the disease course evolution.

1. Introduction

Multiple Sclerosis (MS) is a neurodegenerative and chronic disease of the central nervous system characterized by damages to the myelin sheaths, resulting in a wide range of symptoms, such as fatigue, numbness, visual disturbances, bladder problems, mobility issues and cognitive deficits.

People with MS (PwMS) are mainly classified according to their disease course: relapsing-remitting (RR), secondary-progressive (SP), primary-progressive (PP) and progressive-relapsing (PR) (Giovannoni et al., 2016). Neurological disability in RR patients is mainly due to the development of multifocal inflammatory lesions and it results in relapses, that are attacks of neurological worsening, followed by partial or complete recovery. Disability accrues predominantly in progressive courses (SP, PP, PR) that are more characterized from diffuse immune mechanisms and neurodegeneration. An estimated 15% of PwMS have a PP or PR course at the onset, the remaining 85% is diagnosed with a RR course. About

80% of *RR* patients develop *SP* course within 15–20 years if untreated, or if the adopted pharmacological and rehabilitative protocols are not continuously adjusted according to the evolution of the disease (Scalfari et al., 2014).

For this reason, the prediction of the transition from *RR* to *SP* is one of the most important methodological gaps that MS researchers are currently addressing. The availability of a statistical model able to predict disease worsening is one of the major unmet needs that could significantly improve timeliness, personalization and, consequently, the efficacy of the treatments. Nowadays, there are no clear clinical, imaging, immunologic or pathologic criteria to foresee the transition from *RR* to *SP* (Lublin et al., 2014). Several clinical factors relating to possible *SP* course predictors have been identified (Bergamaschi et al., 2015; Dickens et al., 2014). However, as showed by Vukusic and Confavreux (2003), studies investigating on prognostic factors for MS course evolution generally suffer from two shortcomings: they report a high proportion of *RR* patients not monitored enough to reach progressive course and they lead, to some extent, to contradictory results. Currently, MS research mainly focuses on developing and assessing drugs and rehabilitative protocols for *RR* patients disregarding progressive courses.

In the recent past, researchers explored the potential role of Patient-Centered Outcomes (PCOs) to follow the progression of neurodegenerative diseases and to take timely healthcare decisions (Black, 2013). PCOs comprise self- and physician-administered tests, questionnaires and clinical scales consisting of either ordinal or categorical scaled answers. As opposed to stressful, not frequently repeatable and expensive clinical exams, like magnetic resonance imaging or blood tests, PCOs are patient-friendly and low-cost measures that could allow to investigate the individual changes and disease impact on several aspects such as physical, cognitive, psychological, social and well-being domains (Fiorini et al., 2015). To date, PCOs are extensively used to assess general health status, to support diagnosis and monitor progress of disease and to quantify the patients' perception of the effectiveness of a given therapy or procedure (Nelson et al., 2015). Nevertheless, it is still unclear which are the most informative PCOs and, contextually, whether they can be used as *predictors* for disease evolution.

In our study, we propose a machine learning approach that, leveraging on *PCO* data, aims at predicting the temporal evolution of MS disease course providing insights on the most appropriate use of PCOs. We resort to a vast category of predictive models, ranging from sparse regularization to ensemble and deep learning methods. These models are widely adopted in the biomedical context as they benefit from good generalization properties as well as they allow to address regression and classification problems within the same statistical and computational framework (LeCun et al., 2015; Qi, 2012; Nowak et al., 2011; Teramoto et al., 2009; Zou and Hastie, 2005).

The remainder of the paper is organized as follows: in Section 2 we present an overview of the different machine learning methods explored; in Section 3 we describe the collected *PCO* data set; in Section 4 we present a thorough description of the adopted experimental design and we describe the proposed predictive model for the evolution of the disease; the obtained results are presented in Section 5 and, finally, our conclusions are drawn in Section 6.

2. Machine learning background

We consider a data set composed of T collections of n_t input-output pairs $\{\boldsymbol{x}_i^t, y_i^t\}_{i=1}^{n_t}$ for $t = 1, \ldots, T$, where $\boldsymbol{x}_i^t \in \mathbb{R}^d$ and $y_i^t \in \{\pm 1\}$ are a d-dimensional representation and a binary label corresponding to the MS disease course diagnosed for the i-th patient at time point t, respectively. For convenience, we identify *SP* as the positive class. The representation of each patient at a fixed time point consists of a d-dimensional vector carrying the answers to the set of PCOs described in Section 3. For binary classification problems, the data set is organized in a data matrix $X \in \mathbb{R}^{N \times d}$, where $N = \sum_{t=1}^{T} n_t$, and a label vector $\boldsymbol{y} \in \{\pm 1\}^N$. Conversely, in multiple-output regression problems we refer to an input matrix $X \in \mathbb{R}^{N \times d}$ and an output matrix $Y \in \mathbb{R}^{N \times k}$ for k tasks, with $1 < k \leq d$ in our case.

2.1 Regularization methods for binary classification

Regularization methods formulate the learning task as the minimization problem of Equation (1), where the *loss function* $V(y_i, f(\boldsymbol{x}_i))$ is a data fidelity term, the *regularization penalty* $R(f)$ introduces additional information used to solve the problem and the parameter λ controls the trade-off between the two terms.

$$\min_f \frac{1}{n} \sum_{i=1}^{n} V(y_i, f(\boldsymbol{x}_i)) + \lambda R(f) \tag{1}$$

Different choices for $V(y_i, f(\boldsymbol{x}_i))$ and $R(f)$ lead to different learning machines. In linear models $f(\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{w}$, a weight vector $\boldsymbol{w} \in \mathbb{R}^d$ is learned from the training data. In particular, linear models are said to be *sparse* when they have only a limited number of nonzero weights w_i for $i = 1, \ldots, d$.

In this work, we take advantage of: *Sparse Logistic Regression* (SLR) (Hastie et al., 2009), *Elastic-Net* (EN) (De Mol et al., 2009; Zou and Hastie, 2005) and *Support Vector Machine* (SVM) (Evgeniou et al., 2000). In Table 1 we illustrate the differences among the three methods in terms of loss function and regularization penalty. The sparsity enforcing penalties of SLR and EN allow to use such models as embedded variable selection methods (Guyon and Elisseeff, 2003). Conversely, to achieve a sparse model with SVM a Recursive Feature Elimination (RFE) wrapper scheme (Guyon et al., 2002) can be used.

Table 1: Overview of the adopted classification loss functions and regularization penalties.

	$V(y_i, f(\boldsymbol{x}_i))$	$R(f)$
EN	$(1 - y_i \boldsymbol{x}_i^T \boldsymbol{w})^2$	$\frac{1}{2}(1-\alpha) \left\|\boldsymbol{w}\right\|_2^2 + \alpha \left\|\boldsymbol{w}\right\|_1$
SLR	$\log(1 + e^{-y_i \boldsymbol{x}_i^T \boldsymbol{w}})$	$\left\|\boldsymbol{w}\right\|_1$
SVM	$\lvert 1 - y_i \boldsymbol{x}_i^T \boldsymbol{w} \rvert_+$	$\left\|\boldsymbol{w}\right\|_2^2$

2.2 Ensemble methods for binary classification

The key idea behind ensemble methods is to build a prediction model by aggregating a collection of multiple *base learners* that are trained to solve the same problem (Zhou, 2012).

Random Forests (RF) is a popular ensemble method that achieves robust predictions by aggregating the estimates of a potentially large number of decision trees constructed from bootstrap samples (Breiman, 2001). Each base learner of a RF is de-correlated from the others as it is built considering only a randomly sampled number of variables $d_b < d$. RF models can be used to perform variable ranking as they embed a measure of relative variable importance (Hastie et al., 2009). In this work, variable selection with RF is achieved by using a RFE schema, as in (Granitto et al., 2006).

Gradient boosting (GB) is a different ensemble learning method based on decision trees (Friedman, 2001). The key idea behind GB is that, under some general hypothesis on the cost function, boosting can be seen as an iterative gradient method for numerical optimization (Hastie et al., 2009). In GB a measure of variable importance can be estimated as in RF. In this work variable selection with GB is achieved by using a RFE schema.

2.3 Regularization methods for multiple-output regression

Regularization methods can also be used to simultaneously learn multiple (k) tasks. Focusing on linear models $f(\boldsymbol{x}_i) = \boldsymbol{x}_i^T W$ the learned weights are $W \in \mathbb{R}^{d \times k}$.

In this study, linear multiple-output regression is achieved by penalizing the loss function $V(Y, f(X)) = \|XW - Y\|_F^2$ with two different regularization penalties: the nuclear norm $\|W\|_*$ and a combination of the the Frobenius norm and the mixed $L_{2,1}$-norm $\frac{1}{2}(1-\alpha)\|W\|_F^2 + \alpha\|W\|_{2,1}$. The learning problem with the first choice for $R(f)$ is known as Nuclear Norm Minimization (NNM) and the second is known as multi-task Elastic-Net (MTEN). The definition of the matrix norms above is presented in Table 2, we refer to (Mishra et al., 2013; Swirszcz and Lozano, 2012; Evgeniou and Pontil, 2007) for a thorough description of multiple-output learning methods.

Table 2: Overview of the matrix norms used for multiple-output regression.

Matrix norm	Notation	Definition
Frobenius	$\|A\|_F$	$\sqrt{\text{trace}(A^T A)}$
Nuclear	$\|A\|_*$	$\text{trace}\sqrt{(A^T A)}$
Mixed $L_{2,1}$	$\|A\|_{2,1}$	$\sum_j \|\boldsymbol{a}_j\|_2$

2.4 Multi-layer perceptron for multiple-output regression

Deep learning methods are a broad class of machine learning techniques that, starting from raw data, aim at learning a suitable feature representation and a prediction function, at the same time (LeCun et al., 2015).

In this study, non-linear multi-output regression is achieved by fully connected *Multi-Layer Perceptron* (MLP) architectures which is a popular deep learning method that, composing several non-linear transformations, can simultaneously predict multiple tasks (Chen et al., 2016). We trained MLP models for multiple-output regression optimizing the squared loss and introducing regularization in the solution by weight decay (Min et al., 2016).

3. PCOs data set description

The predictive model presented in this work is based on a *PCO* data set acquired from a cohort of PwMS progressively enrolled within an ongoing funded project. Ethical review committee approval *023REG2014* was obtained for this work.

Each patient is evaluated every four months through the items of the PCOs reported in Table 3 which cover physical, cognitive and psychosocial domains. *PCO* data are intrinsically noisy due to the subjectivity of self-reported measures provided by the patients that can be influenced by personal feelings and opinions. In order to ameliorate this issue, 4 questionnaires out of 10 are administered by medical staff which is trained to keep a homogeneous level of evaluation.

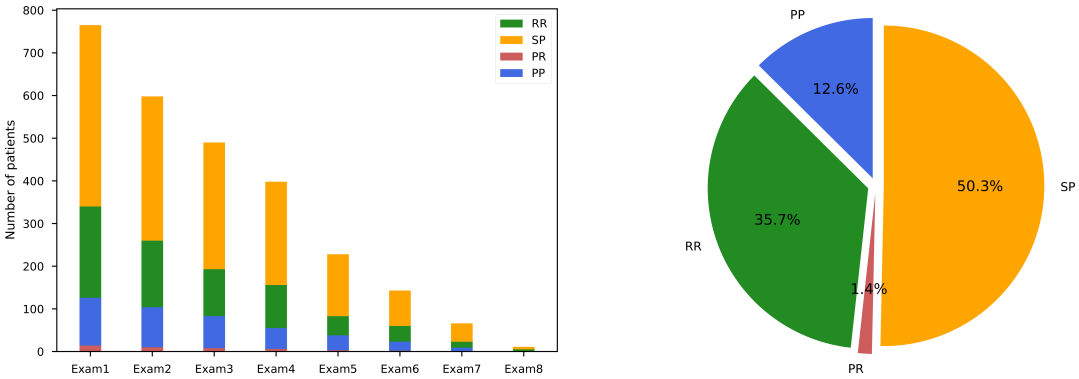
In our analysis we considered all the PCOs reported in Table 3 except EDSS. Such scale is based on a neurological examination and, although usually adopted as an index of the disability level, it focuses mainly on deambulation disability without taking into account other aspects that could impact patient disability, such as upper limb or cognitive functions (Meyer-Moock et al., 2014; Uitdehaag, 2014).

Acronym	**Full name**	**Reference**
$MFIS$	Modified fatigue impact scale	(Flachenecker et al., 2002)
$HADS$	Hospital anxiety and depression scale	(Honarmand and Feinstein, 2009)
$LIFE$	Life satisfaction index	(Franchignoni et al., 1999)
OAB	Overactive bladder questionnaire	(Cardozo et al., 2014)
$EDINB$	Edinburgh handedness inventory	(Oldfield, 1971)
$ABILH$	Hand ability index	(Arnould et al., 2012)
FIM	Functional independence measure	(Granger et al., 1990)
$MOCA$	Montreal cognitive assessment	(Dagenais et al., 2013)
$PASAT$	Paced auditory serial addition task	(Aupperle et al., 2002)
$SDMT$	Symbol digit modality test	(Parmenter et al., 2007)
$EDSS$	Expanded disability status scale	(Kurtzke, 1983)

Table 3: The set of available PCOs. The first 6 are self-reported, while the last 5 are administered by trained medical staff. In our analysis all PCOs were used, with the exception of $EDSS$.

The collected *PCO* data set comprises additional information such as: i) number of relapses in the last four months (NR), ii) educational level expressed in terms of total years of education (EDU), iii) height (H) and iv) weight (W). Each sample of the data set is represented by a vector of $d = 165$ predictors. Moreover, a neurologist assigns to each patient the corresponding disease course. The global distribution of MS types across time points is depicted in Figure 1b.

In this work we focus on predicting MS course evolution of *RR* and *SP* patients, hence the subjects with *PR* and *PP* forms will not be taken into account. We considered all the patients with a minimum of 1 time point (the most recently enrolled) up to $T = 8$ time points for a total of 2699 samples, of which 1220 *RR* and 1579 *SP* (see Figure 1a). As this is an ongoing project, the number of PwMS decreases with time. We expect to fill the gap of samples between *Exam* 1 and *Exam* 8 by the end of the funded study.



(a) (b)

Figure 1: An overview of the *PCO* data set used in this study. The left panel (a) shows a bar chart of the number of MS patients in each disease form at different examinations. The right panel (b) presents a representation of the distribution of the total amount of acquisitions (3137), divided according to the disease form.

4. Problem description

Predicting the MS course evolution can be split in three different related tasks: Current Course Assignment (CCA), PCOs Evolution Prediction (PEP) and Future Course Assignment (FCA). In particular, given the 165-dimensional representation of a patient at a fixed time point $\boldsymbol{x}_i^t$, CCA consists in assigning the corresponding disease course y_i^t. Given the historical representation of a patient $\boldsymbol{x}_i^t$ for $t = 1, \ldots, \tau$, PEP consists in predicting the patient representation $\boldsymbol{x}_i^{\tau+1}$. Finally, FCA consists in foreseeing the MS disease course $y_i^{\tau+1}$ from $\boldsymbol{x}_i^t$ for $t = 1, \ldots, \tau$.

Here, we developed a predictive model that solves these tasks assuming the temporal structure outlined in Figure 2. The CCA problem is translated into a binary classification task and we address it by learning a discriminative function $f(\boldsymbol{x}_i^t) = y_i^t$. The PEP problem is modeled as $g(\boldsymbol{x}_i^t) = \boldsymbol{x}_i^{t+1}$, where $g(\boldsymbol{x})$ is a multiple-output regression function. Once $\hat{f}(\boldsymbol{x})$ and $\hat{g}(\boldsymbol{x})$ are learned by training on historical *PCO* data, the FCA problem is finally solved by the temporal model $\hat{f} \circ \hat{g}(\boldsymbol{x}_i^t) = y_i^{t+1}$. In time-series data analysis, this is known

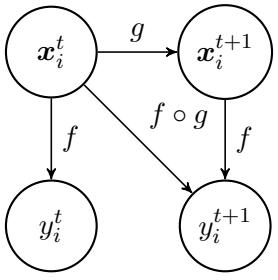


Figure 2: A visual representation of the temporal structure assumed in the collected data. When the two functions f (CCA) and g (PEP) are learned, the FCA model $f \circ g$ is able to predict the evolution of the disease course for future time points y_i^{t+1}.

as *one-step-ahead forecast*. Notably, the FCA model allows to foresee if the patient at the next time point is going to experience a transition from *RR* to *SP*, or not.

4.1 Data preprocessing

Analyzing *PCO* data is challenging from several respects. First, items belonging to different questionnaires are encoded with numerical values in different ranges. To tackle this issue, we opted for a $[0-1]$ scaling of the ordinal answers and a binary one-hot-encoding of the categorical ones. Secondly, as the missing data amount to 1.52% of the entire data set, we resort to the K-nearest neighbor data imputing strategy proposed in (Troyanskaya et al., 2001). To ensure unbiasedness of the results, this preprocessing phase is not performed on the entire data collection, but it is separately evaluated prior to each model fitting process on its cross-validation portion of the training set, as described in the next section.

4.2 Experimental design

We shall discuss separately the experimental designs used to learn $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$.

The CCA model $f(\boldsymbol{x})$ solves a binary classification problem: to each input $\boldsymbol{x}_i^t$ is associated an output y_i^t that encodes the corresponding MS disease course (*RR* or *SP*) with a binary label. We split the data set in three temporal chunks, namely *training*, *validation* and *test* sets, consisting of all samples collected at time points $t = 1, 2, 3$, $t = 4$ and $t = 5, 6, 7, 8$, respectively. Accordingly, we used 1853 samples for training $f(\boldsymbol{x})$, 398 for validation leaving the remaining 448 for test. Five candidate models for $f(\boldsymbol{x})$ are fitted on 20 Monte Carlo (MC) random sampling of the training set each time keeping $\frac{1}{4}$ of the samples aside (Molinaro et al., 2005). For each MC sampling the fitting procedure is performed on the remaining $\frac{3}{4}$ of the samples and it includes an inner parameter optimization via grid-search cross-validation (Hastie et al., 2009). In particular, we require the MS course prediction to be based on a reduced number of variables (see Section 4.3), therefore we enforce sparsity in each candidate model. Leveraging on the MC strategy, we rank the variables according to their selection frequency (Barbieri et al., 2016; Meinshausen and Bühlmann, 2010). Once a variable ranking is achieved for each candidate model, the list of selected variables is identified by thresholding the corresponding ranking with the threshold that maximizes the accuracy on the validation set. Finally, the last training step consists in fitting each candidate model on the union of training and validation sets taking only into account the corresponding reduced subset of selected variables. The final CCA model $\hat{f}(\boldsymbol{x})$ is chosen as the one that performs better on the previously unseen test set in terms of accuracy, precision, recall and F_1 score.

On the other hand, learning the PEP model $g(\boldsymbol{x})$ implies solving a multiple-output regression problem and each input $\boldsymbol{x}_i^t$ is associated with the output vector $\boldsymbol{x}_i^{t+1}$. Therefore, we can only consider samples at time point t with an available follow-up at the next time point $t+1$, which reduces the overall number of available samples. The data set splitting is consistent with the one followed for learning $f(\boldsymbol{x})$, although there is no need for a separate validation set, as learning $g(\boldsymbol{x})$ does not require any variable selection process. We used the samples collected at time points $t = 1, 2, 3, 4$ for training and those at $t = 5, 6, 7, 8$ for test, resulting in 1737 and 254 samples, respectively. The fitting procedure includes an inner parameter optimization via grid-search cross-validation. Each candidate model is a

function $g : \mathbb{R}^{165} \to \mathbb{R}^k$ where k is the number of variables selected by the best CCA model. The final PEP model $\hat{g}(\boldsymbol{x})$ is chosen as the candidate model that performs better on the previously unseen test set in terms of mean absolute error (MAE).

The predictive capability of the FCA model $\hat{f} \circ \hat{g}(\boldsymbol{x})$ is finally evaluated on the test set. The CCA model $\hat{f}(\boldsymbol{x}_i^t)$ predicts the MS course $\hat{y}_i^t$ from the *PCO* data vector $\hat{\boldsymbol{x}}_i^t$ that, in turn, is predicted by the PEP model $\hat{g}(\boldsymbol{x}_i^{t-1})$. We shall notice here that the predictions $\hat{f} \circ \hat{g}(\boldsymbol{x}_i^t) = y_i^{t+1}$ for $t = 8$ are foreseeing possible *RR* to *SP* transitions that are beyond our data observation, hence predictions at the last time point cannot be used to assess the FCA model performance. Therefore, its performance is evaluated only on 220 test samples.

4.3 Learning $f(x)$

We imposed $f(\boldsymbol{x})$ to be sparse. This requirement is helpful from two distinct respects: a) the performance of the predictive model may increase thanks to a reduced effect the course of dimensionality (Hastie et al., 2015) and b) the identification of a reduced subset of meaningful PCOs provides interpretability of the results for the clinicians. In order to achieve such sparse model, we take advantage of two main variable selection strategies: embedded and wrapper methods (Guyon and Elisseeff, 2003). When using embedded methods, we exploited the sparsity inducing penalties of EN to take into account possible correlation between *PCO* variables and of SLR to benefit from the renowned classification capability of the logistic loss function. We applied the RFE wrapper method to two tree-based learning machines (RF and GB) that are capable of capturing nonlinear relationship between input and output and are intrinsically well-suited to deal with categorical/ordinal variables. We also explored the use of RFE with SVM, as in (Guyon et al., 2002).

4.4 Learning $g(x)$

As no prior information on the relationship between PCOs evaluated at different time points was available, to learn $g(\boldsymbol{x})$ we investigated on the use of both linear and nonlinear models.

Concerning the linear models, we explored two different solutions: NNM and MTEN. The first imposes a low-rank prior on the result. The second is a natural multiple-output extension of EN, hence it induces a row-structured sparsity pattern on the solution where collinear variables are more likely to be included in the model together. For nonlinear prediction, we resorted to the state-of-the-art MLP approach.

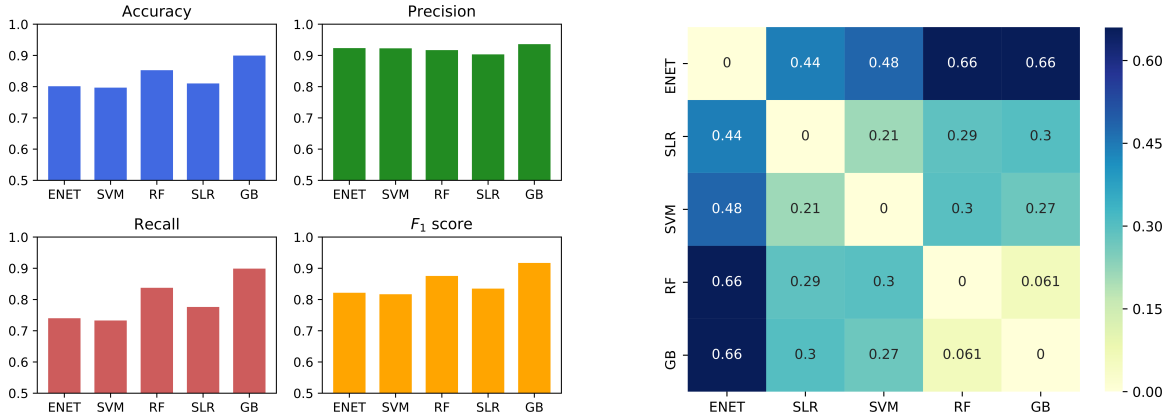
5. Results

We shall discuss separately the results achieved in terms of CCA, PEP and FCA models.

Regarding CCA, the GB method outperforms the other candidate models reaching accuracy 0.900, precision 0.936, recall 0.899 and F_1 score 0.917, as shown in Figure 3a. Therefore we chose it as CCA model $\hat{f}(\boldsymbol{x})$. Insights on the use of PCOs for MS assessment are provided by the sparsity of the CCA model induced by the RFE schema. The 31 selected variables are reported in Table 4. Comparing the full list of *PCO* questionnaires of Table 3 with Table 4, we observe that each *PCO* used in this study is represented at least once, except EDINB, and the most represented is FIM. We also see that, whenever possible, the model tends to select aggregate scores (total and subtotal) rather that single items. This is

consistent with the clinical practice, where neurologists are more likely to assess patient's health status by using the aggregate scores, rather than the single questions. Quite surprisingly, the recent number of relapses is the only additional information not selected by the model. Finally, we note that all the domains that are known to be affected by the disease are well covered: mobility (upper and lower limbs), cognition, emotional, fatigue, bladder and psychosocial. The heatmap in Figure 3b shows the Hamming distance estimated across the list of variables selected by the five CCA candidate models. Interestingly, tree-based methods are more prone to select similar variables with respect to linear methods. As expected, the sparsity induced by the ℓ_1-norm of SLR allows the method to achieve a list of variables similar to the one obtained by SVM-RFE, while the list obtained by ENET includes collinear variables and it is significantly different from the others.

Regarding PEP, MTEN outperforms the other candidate models in terms of MAE ($\text{MAE}_{\text{MTEN}} = 0.095$, $\text{MAE}_{\text{NNM}} = 0.102$, $\text{MAE}_{\text{MLP}} = 0.105$), hence we select it as our PEP model $\hat{g}(\boldsymbol{x})$.



(a) (b)

Figure 3: A visual representation of the results obtained from the CCA model. On the left panel (a) we show the classification performance achieved on the test set by the candidate models. Precision, recall and F_1 score are estimated considering *SP* as the positive class. As GB outperforms the other methods on each performance metric, it is chosen as CCA model. On the right panel (b) a heatmap displays the distance between the lists of variables selected by each model in terms of their hamming distance.

Finally, the FCA model $\hat{f} \circ \hat{g}(\boldsymbol{x})$, obtained by combining MTEN and GB achieves the following performance scores on the 220 test samples: accuracy 0.841, precision 0.900, recall 0.824 and F_1 score 0.860.

Selected item	**Description**
ABILH (TOT)	Sum of all the ABILH subscores
FIM (003)	How much assistance is required for bathing
FIM (009)	How much assistance is required for bed to chair transfer
FIM (010)	How much assistance is required for toilet transfer
FIM (011)	How much assistance is required for shower transfer
FIM (012)	How much assistance is required for locomotion (ambulatory)

FIM (014)	How much assistance is required for locomotion (wheelchair)
FIM (SUB3)	FIM subtotal measuring global sphincter control
FIM (SUB4)	FIM subtotal measuring global personal care
FIM (SUB5)	FIM subtotal measuring global locomotion
FIM (SUB6)	FIM subtotal measuring global mobility
FIM (TOT)	FIM total score
HADS (SUB1)	HADS subtotal measuring global level of anxiety
HADS (SUB2)	HADS subtotal measuring global level of depression
HADS (TOT)	HADS total score
H	Height of the individual in cm
EDU	Total years of formal education
LIFE (TOT)	LIFE total score
MFIS (002)	I have had difficulty paying attention for long periods of time
MFIS (SUB1)	MFIS subtotal measuring global cognitive level
MFIS (SUB2)	MFIS subtotal measuring global physical level
MFIS (SUB3)	MFIS subtotal measuring global psychosocial level
MFIS (TOT)	MFIS total score
MOCA (001)	MOCA visuoconstructional skill test
MOCA (009)	MOCA memory test
MOCA (SUB1)	Sum of all the MOCA subscores
MOCA (TOT)	MOCA score corrected for individuals with less than 12 years of formal education
OAB (TOT)	OAB total score
PASAT	PASAT score
SDMT	SDMT score
W	Weight of the individual in kg

Table 4: The list of PCOs items selected by GB with RFE.

6. Conclusion

In this work we proposed a novel temporal model based on patient-centered outcomes and machine learning for disease form prediction in multiple sclerosis. In particular, we address the tasks of current course assignment, PCOs evolution prediction and future course assignment. The model is built on a collection of PCOs acquired on a cohort of individuals enrolled in an ongoing funded study (*DETECT-MS PRO*). PCOs data are typically used to corroborate evidence provided by quantitative exams, in our case the absence of clear MS disease form predictors makes the information extracted from PCOs data the only available resource. The proposed temporal model was able to correctly assign the current MS form and to foresee future ones with accuracy of 90.0% and 84.1%, respectively. This demonstrates that PCOs can effectively be used as MS disease course predictor. In the next future, we plan to further investigate on the predictive capabilities of the proposed model with longer temporal horizons and to compare it with different approaches, such as probabilistic graphical models. Given the achieved promising results, the proposed model is soon going to be validated in clinical practice, where it will assist the clinicians involved in this study to foresee possible disease course transition and to take important decisions concerning treatment and therapies that can substantially improve the quality of life of their patients. In the era of precision medicine, the problem of predicting MS course evolution still relies on stressful exams and clinical judgement. To the best of our knowledge, this is the first attempt to solve this delicate task leveraging on patient-friendly measures and machine learning.

Acknowledgments

Supported by FISM - Fondazione Italiana Sclerosi Multipla - cod. 2015/R/03.

References

Carlyne Arnould, Laure Vandervelde, Charles Sèbiyo Batcho, Massimo Penta, and Jean-Louis Thonnard. Can manual ability be measured with a generic abilhand scale? a cross-sectional study conducted on six diagnostic groups. *BMJ open*, 2(6):e001807, 2012.

Robin L Aupperle, William W Beatty, Fàtima de N.A.P. Shelton, and Samuel T Gontkovsky. Three screening batteries to detect cognitive impairment in multiple sclerosis. *Multiple Sclerosis*, 8(5):382–389, 2002.

Matteo Barbieri, Samuele Fiorini, Federico Tomasi, and Annalisa Barla. PALLADIO: a parallel framework for robust variable selection in high-dimensional data. *PyHPC2016 conference, IEEE proceedings*, 2016.

Roberto Bergamaschi, Cristina Montomoli, Giulia Mallucci, Alessandra Lugaresi, Guillermo Izquierdo, Francois Grand'Maison, Pierre Duquette, Vahid Shaygannejad, Raed Alroughani, Pierre Grammond, et al. Bremso: A simple score to predict early the natural course of multiple sclerosis. *European journal of neurology*, 22(6):981–989, 2015.

Nick Black. Patient reported outcome measures could help transform healthcare. *BMJ (Clinical research ed)*, 346:f167, 2013.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Linda Cardozo, David Staskin, Brooke Currie, Ingela Wiklund, Denise Globe, Manuel Signori, Roger Dmochowski, Scott MacDiarmid, Victor W Nitti, and Karen Noblett. Validation of a bladder symptom screening tool in women with incontinence due to overactive bladder. *International urogynecology journal*, 25(12):1655–1663, 2014.

Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 2016.

Emmanuelle Dagenais, Isabelle Rouleau, Mélanie Demers, Céline Jobin, Élaine Roger, Laury Chamelian, and Pierre Duquette. Value of the moca test as a screening instrument in multiple sclerosis. *The Canadian Journal of Neurological Sciences*, 40(03):410–415, 2013.

Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco. Elastic net regularization in learning theory. *Journal of Complexity*, 2009.

Alex M Dickens, James R Larkin, Julian L Griffin, Ana Cavey, Lucy Matthews, Martin R Turner, Gordon K Wilcock, Benjamin G Davis, Timothy DW Claridge, Jacqueline Palace, et al. A type 2 biomarker separates relapsing-remitting from secondary progressive multiple sclerosis. *Neurology*, 83(17):1492–1499, 2014.

Theodoros Evgeniou and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

Samuele Fiorini, Alessandro Verri, Andrea Tacchino, Michela Ponzio, Giampaolo Brichetto, and Annalisa Barla. A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 4443–4446. IEEE, 2015.

Peter Flachenecker, Tania Kümpfel, B Kallmann, M Gottschalk, O Grauer, P Rieckmann, C Trenkwalder, and KV Toyka. Fatigue in multiple sclerosis: a comparison of different rating scales and correlation to clinical parameters. *Multiple sclerosis*, 8(6):523–526, 2002.

Franco Franchignoni, Luigi Tesio, Marcella Ottonello, and Emilio Benevolo. Life satisfaction index: Italian version and validation of a short form1. *American journal of physical medicine & rehabilitation*, 78(6):509–515, 1999.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Gavin Giovannoni, Helmut Butzkueven, Suhayl Dhib-Jalbut, Jeremy Hobart, Gisela Kobelt, George Pepper, Maria Pia Sormani, Christoph Thalheim, Anthony Traboulsee, and Timothy Vollmer. Brain health: time matters in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, 9:S5–S48, 2016.

CV Granger, AC Cotter, BB Hamilton, RC Fiedler, and MM Hens. Functional assessment scales: a study of persons with multiple sclerosis. *Archives of physical medicine and rehabilitation*, 71(11):870–875, 1990.

Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, 2006.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

Kimia Honarmand and Anthony Feinstein. Validation of the hospital anxiety and depression scale for use with multiple sclerosis patients. *Multiple Sclerosis*, 2009.

John F Kurtzke. Rating neurologic impairment in multiple sclerosis an expanded disability status scale (edss). *Neurology*, 33(11):1444–1444, 1983.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015.

Fred D Lublin, Stephen C Reingold, Jeffrey A Cohen, Gary R Cutter, Per Soelberg Sørensen, Alan J Thompson, Jerry S Wolinsky, Laura J Balcer, Brenda Banwell, Frederik Barkhof, et al. Defining the clinical course of multiple sclerosis the 2013 revisions. *Neurology*, 83 (3):278–286, 2014.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Sandra Meyer-Moock, You-Shan Feng, Mathias Maeurer, Franz-Werner Dippel, and Thomas Kohlmann. Systematic literature review and validity evaluation of the expanded disability status scale (edss) and the multiple sclerosis functional composite (msfc) in patients with multiple sclerosis. *BMC neurology*, 14(1):58, 2014.

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *arXiv preprint arXiv:1603.06430*, 2016.

Bamdev Mishra, Gilles Meyer, Francis Bach, and Rodolphe Sepulchre. Low-rank optimization with trace norm penalty. *SIAM Journal on Optimization*, 23(4):2124–2149, 2013.

Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 2005.

Eugene C Nelson, Elena Eftimovska, Cristin Lind, Andreas Hager, John H Wasson, and Staffan Lindblad. Patient reported outcome measures in practice. *Bmj*, 350:g7818, 2015.

Gen Nowak, Trevor Hastie, Jonathan R Pollack, and Robert Tibshirani. A fused lasso latent feature model for analyzing multi-sample acgh data. *Biostatistics*, page kxr012, 2011.

Richard C Oldfield. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

BA Parmenter, B Weinstock-Guttman, N Garg, F Munschauer, and R HB Benedict. Screening for cognitive impairment in multiple sclerosis using the symbol digit modalities test. *Multiple Sclerosis*, 13(1):52–57, 2007.

Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

Antonio Scalfari, Anneke Neuhaus, Martin Daumer, Paolo Antonio Muraro, and George Cornell Ebers. Onset of secondary progressive phase and long-term evolution of multiple sclerosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 85(1):67–75, 2014.

Grzegorz Swirszcz and Aurelie C Lozano. Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 361–368, 2012.

Reiji Teramoto et al. Balanced gradient boosting from imbalanced data for clinical outcome prediction. *Statistical applications in genetics and molecular biology*, 8(1):1–19, 2009.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

BM Uitdehaag. Clinical outcome measures in multiple sclerosis. *Handb Clin Neurol*, 122: 393–404, 2014.

Sandra Vukusic and Christian Confavreux. Prognostic factors for progression of disability in the secondary progressive phase of multiple sclerosis. *Journal of the neurological sciences*, 206(2):135–137, 2003.

Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.