Development of predictive models for short-term prediction of disability progression in multiple sclerosis

E. Tavazzi[1], D. Albertini[1], and M. Vettoretti[1*], for MSOAC[†]
[1] *Department of Information Engineering, University of Padova, Padova, Italy*
[*] *Corresponding author: martina.vettoretti@unipd.it.*

***Abstract*—Multiple Sclerosis (MS) is an autoimmune degenerative disease of the central nervous system, in which chronic inflammation leads to demyelination with transient or permanent axon damage. Symptoms of MS include problems with vision, movement, sensation and balance, which can be intermittent or progressively increasing over time until bringing to permanent disability. Predictive models of MS disability progression can be very useful to support the clinician in choosing the best care for each patient.**

The aim of this work is to develop predictive models of short-term MS disability progression. Data are part of the Multiple Sclerosis Outcome Assessments Consortium (MSOAC) Placebo database, which includes longitudinal demographic and clinical data of 2465 MS patients who were enrolled in the control arm of different MS clinical trials. Variables collected in the first visit were used to predict a binary outcome of disability progression at 6 months and 18 months from the baseline, using a logistic regression model. Disability progression was defined as a 1.5 increase in the Expanded Disability Status Scale (EDSS) value compared to the baseline time. 20 input variables were considered in each model, including demographics, medical history, functional tests, questionnaires, and MS phenotype. Preprocessed data were split into a training and a test set with an 80%-20% proportion. Logistic regression models were trained on the training set, using over-/undersampling techniques for balancing the classes. The identified models were tested on the test set by assessing the area under the receiver operating characteristic curve (AUC).

Prediction performance on the test set was satisfactory, although not optimal, with AUC equal to 0.74 at 6 months and 0.71 at 18 months. These prediction performances are comparable with results obtained by other literature studies on smaller cohorts.

Future developments of this work include the use of other machine learning techniques for model training, the application of feature selection and variable ranking techniques, the incorporation of new variables (e.g., imaging variables), and the external validation of the models on new populations.

***Keywords*—Predictive Model, Multiple Sclerosis, Expanded Disability Status Scale, Disability Progression**

I. Introduction

Multiple Sclerosis (MS) is a chronic autoimmune and inflammatory disease involving the central nervous system, characterised by demyelination and transient or permanent damage to axons [1]. MS is a multifactorial disease whose causes are still not fully determined. Probable underlying causes include genetics, environmental factors, as well as viral infections [2]. MS causes a variety of physical and cognitive symptoms. The course of the disease can consist of isolated recurring attacks, named relapses, that characterise the relapsing-remitting MS (RRMS) phenotype, or of a progressive decline of the subject's functionalities, that can either occur since the first phases of the disease (Primary Progressive Multiple Sclerosis, PPMS) or in a second phase only (Secondary Progressive Multiple Sclerosis, SPMS). The care for MS mainly consists of pharmacological treatments administered to delay disease progression or mitigate the symptoms.

In this context, the employment of prognostic predictive models would lead to many advantages, including promoting the understanding of the mechanisms of MS, anticipating treatments for patients in need of preventive therapies, thereby improving the risk/benefit ratio and the quality of life of these individuals, as well as informing patients and caregivers about the future evolution of the disease. Several literature studies have investigated the prediction of progression in terms of different outcomes, such as the values of functional tests performed by the patients or a related score [3], [4], the increase of fatigue [5], the shift from one MS phenotype to another [6], [7], or the time/chance of occurrence of a relapse [8], [9]. To this end, various classical machine learning methodologies have been used, adopting regression, classification, or survival analysis approaches.

In this work, we focused on modelling the progression of MS in terms of the worsening of the patient's disabilities. We considered as the outcome the Expanded Disability Status Scale (EDSS) [10], a score that is used in clinical practice for monitoring the status and capacities of MS patients and whose variation over time is typically used to define an exacerbation of the disease. The EDSS ranges from 0 (normal status) to 10 (death due to MS), with increasing values as the disability progresses. This score incorporates the status of eight functional systems (pyramidal, cerebellar, brain stem, sensory, bowel and bladder, visual, mental, and other functions), which together express all possible neurological alterations induced by the disease. After defining a threshold on EDSS variations and a prediction horizon, i.e., a temporal time-window in

[†]Data used in the preparation of this article were obtained from the Multiple Sclerosis Outcome Assessments Consortium (MSOAC). As such, the investigators within MSOAC contributed to the design and implementation of the MSOAC Placebo database and/or provided placebo data, but did not participate in the analysis of the data or the writing of this report.

which the subject is observed, each subject can be associated with a binary outcome (progression of disability occurred or not), and the prediction can be approached as a classification task.

This approach mimics the ones proposed in the literature by Zhao et al. [11], [12], by Law et al. [3], by Tommasin et al. [13], and by Yperman et al. [14]. In that literature studies, however, we observed that the cardinality of the cohorts used for model development was often limited to only a few hundred patients, which clashes with the requirements of machine learning to have a sufficiently large number of instances to learn models appropriately. To overcome this issue, in this work we employed the Multiple Sclerosis Outcome Assessments Consortium (MSOAC) Placebo database [15], [16], which includes demographic and clinical data of 2465 MS patients who were involved in clinical trials of MS treatments as the control groups. Starting from the first recorded visit (baseline), we identified two prediction horizons, namely 6 and 18 months after the baseline, computed the outcomes, and developed a logistic regression model for each setting using as input the variables collected during the first trial visit. We then assessed the prediction performance of the model on an independent test set.

II. Materials and Methods

A. Dataset and preprocessing

The MSOAC Placebo Database includes fully anonymized longitudinal data of 2465 MS patients from 9 clinical trials [15], [16]. The database contains data on demographics, medical history, performance outcome measures (e.g., Timed 25 Foot Walk (T25FW), 9-Hole Peg Test (NHPT), Paced Auditory Serial Addition Test (PASAT)), clinician-reported outcome measures (e.g., EDSS, Kurtzke Functional Systems Scores (KFSS)), patient-reported outcome measures (e.g., 36-Item Short Form Health Survey (SF-36)), relapse information, and MS phenotype (e.g., relapsing-remitting), for a total of 104 longitudinally-collected variables. The MSOAC database does not include imaging data. Due to the nature of the dataset, enrollment time and follow-up duration may vary for subjects participating in different clinical trials.

We extracted the information collected at the first visit of each patient to be used as input for the predictive models and preprocessed it as follows. After removing the data that referred to an under-represented MS phenotype (Progressive-Relapsing MS, 2 patients only), we filtered the variables to remove the ones with too much missing information. By excluding the variables with more than 30% of missing values, the total number of features was reduced to 23. We then filtered out three more boolean features – namely the one indicating the need for more than 2 attempts to perform the T25FW test, the NHPT with the dominant hand, or with the non-dominant hand – that were excluded since the occurrence of positive cases was extremely limited. This brought the final number of features to 20. Table I reports a description of the included variables, with the percentage of their missing values (fourth column), the mean and interquartile range (IQR) of numerical variables or the levels of categorical variables (fifth column).

B. Model development pipeline

Data collected at the first visit of each subject, referred to as baseline, were used to predict subject's disease progression at two prediction horizons (PH), i.e., at 6 and 18 months after the time of the baseline visit (t_B). For each subject, the EDSS value at baseline, $EDSS_{t_B}$, was defined as the first available EDSS value. The EDSS at PH, $EDSS_{t_B+PH}$, was defined as the EDSS value recorded at $t_B + PH$. Subjects with no EDSS value recorded at $t_B + PH$ were excluded from the analysis. For each PH value, a binary outcome of disability progression was defined as follows. A subject was provided label "1", i.e., progression of disability, if $EDSS_{t_B+PH} - EDSS_{t_B} \geq 1.5$ in accordance with the work by Zhao et al. [12]. Conversely, a subject was assigned label "0", i.e., no progression of disability, if $EDSS_{t_B+PH} - EDSS_{t_B} < 1.5$.

A different predictive model was developed for each PH using logistic regression as a binary classification model. Labelled data were randomly split into a training set, with 80% of the subjects, and a test set, with the remaining 20% of the subjects. The training-test split was stratified by model outcome in order to maintain the same proportion of classes in the two sets. Missing values in the input variables were imputed using the Multiple Imputation by Chained Equations (MICE) method [17]. The parameters of the imputation algorithm were learned from the training set, and used to impute both the training and the test set. Then, categorical variables were encoded using dummy variables. Finally, numerical variables were scaled in the range 0-1, by subtracting to each variable its minimum value in the training set, and then dividing the result by the range of the variable values in the training set. In order to manage the strong class unbalance, two strategies were adopted for balancing the classes on the training set. The first one consisted in oversampling the less frequent class, until the two classes are balanced ("Oversampling"). The second one consisted in oversampling the less frequent class, and undersampling the more frequent class, until the two classes are balanced ("Both Over-/Under-sampling").

A logistic regression model was trained on the balanced training set. The logistic regression equation is:

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_n \cdot X_n) \qquad (1)$$

where P is the probability of the class "1", X_i, i = 1,..., n are the input variables, β_i, i = 1,...,n are the model parameters associated with the input variables, and β_0 is the intercept parameter. Model parameters were estimated on the training set by maximum likelihood. Variables significantly associated with the outcome were identified by the Wald test (significance level 5%).

In total 4 models were developed (2 PH values x 2 strategies for balancing the training set).

TABLE I: BASELINE VARIABLES SELECTED AS INPUT FOR THE PREDICTIVE MODEL AFTER DATA PREPROCESSING.

Domain	Name	Description	NA (%)	Mean [IQR] / Levels
Demographics	AGE	Age at first visit (in years)	3.37	41.77 [18, 72]
	SEX	Sex	0	Male / Female
Findings about Medical History	P1Y_RELAPSES	Number of relapses experienced in the year previous to the first visit	27.75	1.376 [0, 7]
	T25_FW_first	First T25FW test performed: time taken to cover a straight stretch of 25 feet (in seconds)	0.81	9.217 [0.7, 218.2]
	T25_FW_second	Second T25FW test performed: time taken to cover a straight stretch of 25 feet (in seconds)	0.85	9.095 [0.7, 232.1]
	PASAT_tc_3s	PASAT: number of correct answers (out of 60) presented every 3 seconds each	3.41	47.24 [0, 60]
Functional Tests	NHPT01_DH1	First NHPT performed with the dominant hand: time taken to complete the test (in seconds)	3.41	25.1 [1, 300]
	NHPT01_DH2	Second NHPT performed with the dominant hand: time taken to complete the test (in seconds)	3.45	23.88 [8.5, 300]
	NHPT01_NDH1	First NHPT performed with the non-dominant hand: time taken to complete the test (in seconds)	3.77	26.87 [1.4, 312.7]
	NHPT01_NDH2	Second NHPT performed with the non-dominant hand: time taken to complete the test (in seconds)	3.85	26.16 [1.2, 356.7]
	EDSS	EDSS score value (in the 1-10 range)	3.29	3.406 [0, 6.5]
	KFSS101	KFSS score value on pyramidal functions (in the 0-9 range)	15.62	1.853 [0, 5]
	KFSS102	KFSS score value on cerebellar functions (in the 0-9 range)	13.35	1.29 [0, 9]
	KFSS103	KFSS score value on brain stem functions (in the 0-9 range)	15.62	0.6351 [0, 4]
Questionnaires	KFSS104	KFSS score value on sensory functions (in the 0-9 range)	15.62	1.178 [0, 5]
	KFSS105	KFSS score value on bowel and bladder functions (in the 0-9 range)	15.62	0.8322 [0, 6]
	KFSS106	KFSS score value on visual functions (in the 0-9 range)	15.62	0.7543 [0, 6]
	KFSS107	KFSS score value on cerebral (or mental) functions (in the 0-9 range)	15.62	0.5827 [0, 3]
	DH	Dominant hand	3.25	Left / Right
Medical History	PHENOTYPE	MS phenotype at first visit	0	RRMS / SPMS / PPMS

C. Performance metrics

Each model was applied to the test set to assess its performance. As performance metrics, we considered the receiver-operating-characteristic (ROC) curve, i.e., the plot of sensitivity vs. 1-specificity for different classification thresholds on the predicted probability of class "progression of disability" (P in eq. 1), and the area under the ROC curve (AUC). The AUC varies between 0 and 1, with 0.5 corresponding to the model that randomly assigns the classes, 1 corresponding to the model that perfectly classifies the data, and 0 to the model that perfectly divides the two classes but switches them.

III. Results

There were a total of 2232 subjects with a definite outcome at 6 months after baseline, of whom 103 had a progression of disability (i.e., EDSS worsening $\geq$ 1.5) and the other 2129 have no progression of disability (i.e., EDSS worsening $<$ 1.5). At the 18-month PH, 1605 subjects have a definite outcome, of whom 164 are in the "progression of disability" class and 1441 in the "no progression of disability" class. After splitting the data, the training and the test sets for the prediction of disability progression at 6 months include 1786 and 446 subjects, respectively. For the 18-month PH, the number of subjects in the training and the test sets is 1284 and 321, respectively.

The AUC values obtained on the test set are reported in Table II. At 6 months, the models achieve satisfactory, though not optimal, prediction performance, with the best results obtained by the model trained with oversampling of the training set. The ROC curve for this model is shown in Fig. 1. At 18 months, the model prediction performance is slightly lower than those at 6 months. At 18 months, the technique that performs both over- and undersampling of the training set achieves a slightly higher prediction performance. The ROC curve for this model is shown in Fig. 2.

In Table III we report for each final model the list of variables that resulted significantly associated with the outcome according to the Wald test (significance level 5%) and the corresponding p-values.

TABLE II: AUC VALUES ON THE TEST SET

Prediction horizon	*Sampling strategy*	*AUC value*
6 months	Oversampling	0.7404
6 months	Both over-/undersampling	0.7284
18 months	Oversampling	0.6969
18 months	Both over-/undersampling	0.7101

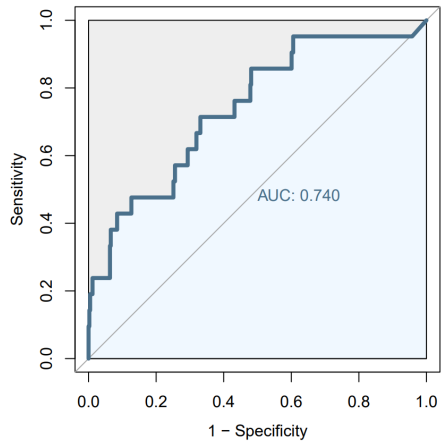

Fig. 1: ROC curve on the test set of the logistic regression model that predicts disease progression at 6 months, using the oversampling technique for balancing classes on the training set.

TABLE III: P-VALUES OF THE VARIABLES ASSOCIATED WITH MODEL OUTCOME ACCORDING TO THE WALD TEST IN AT LEAST ONE MODEL

Variable	*Model PH=6*	*Model PH=18*
AGE	4.95e-15	
SEX	0.033533	
P1Y_RELAPSES	< 2e-16	0.026148
T25FW_first	7.97e-05	
T25FW_second	9.14e-06	
PASAT_tc_3s	< 2e-16	
NHPT01_DH1	4.72e-06	0.012367
NHPT01_DH2	0.001201	
EDSS	8.70e-12	7.33e-11
KFSS101		0.000386
KFSS105		0.006603
DH		0.030788
PHENOTYPE	0.000287	2.11e-05

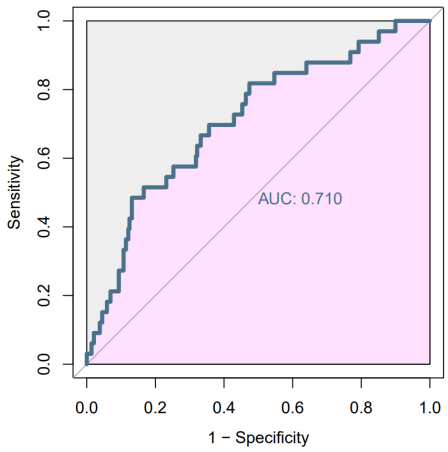

Fig. 2: ROC curve on the test set of the logistic regression model that predicts disease progression at 18 months, using both over-/undersampling techniques for balancing classes on the training set.

IV. Conclusion

In this work, we used the data of the MSOAC Placebo Database to develop logistic regression models of short-term disability progression, defined as an increase of the EDSS score larger than 1.5, with a PH of 6 months and 18 months. The developed models achieved satisfactory prediction performance on an independent test set, with AUC values of 0.74 at 6 months and 0.71 at 18 months. The slight deterioration of prediction performance at 18 months is somehow expected because, given the large variability among disease progression profiles observed in MS, the larger the PH the more difficult is to predict with confidence the progression of disability. These prediction performances are comparable with results obtained by other literature studies on smaller cohorts (AUC=0.62 at 6 months in Law et al. [3] using a cohort of 485 patients; AUC=0.79 at 24 months in Tommasin et al. [13] using a cohort of 163 patients; AUC=0.75 at 24 months in Yperman et al. [14] using a cohort of 642 patients), confirming the feasibility of MS progression prediction.

By comparing the variables most significant for outcome prediction at different PH (Table III), we can observe that some variables are significantly associated to progression of disability at both 6 and 18 months (i.e., number of relapses in the last year, EDSS, NHPT01_DH1, and MS phenotype). Interestingly, the role of other variables varies with the PH, with demographic variables (i.e., age and sex) being not significantly associated to progression of disability at 18 month.

The prediction performance is still not optimal, but several margins for improvement exist. Although the dataset used in this study is larger than those used by studies in the literature, indeed, this dataset poses some major challenges, such as significant class imbalance, absence of imaging data, short follow-up period, and possibly different enrollment times. Class imbalance can significantly affect model training, although in this study we mitigated its effect by using over/undersampling techniques. Magnetic resonance imaging (MRI) of the central nervous system is the main diagnostic test to quantify the effects of demyelination, and it is routinely performed to monitor MS progression. Therefore, the inclusion of MRI-derived variables in the models could bring some important information to improve the prediction of disease progression. Finally, datasets with longer follow-up periods could allow the development of predictive models of long-term disease progression.

Future developments of this work include the use of other machine learning techniques for model development (e.g., support vector machines or random forest), application of feature selection and variable ranking techniques, incorporation of new variables (e.g., MRI variables if available), and external validation of the models on new patient populations.

Acknowledgement

This work was supported by MIUR (Italian Ministry for Education) under the initiative "Departments of Excellence" (Law 232/2016), the Department of Information Engineering of the University of Padova, Padova, Italy ("Dotazione Ordinaria per la Ricerca" 2019, and "Research Grant (type B) – junior initiative" 2021.)

References

[1] Compston, A. & Coles, A. Multiple sclerosis. *Lancet* **372**, 1502–17 (2008).
[2] Kamm, C. P. *et al.* Multiple sclerosis: current knowledge and future outlook. *European neurology* **72**, 132–141 (2014).
[3] Law, M. T. *et al.* Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression. *Multiple Sclerosis Journal–Experimental, Translational and Clinical* **5**, 2055217319885983 (2019).
[4] Tsagkas, C. *et al.* Central nervous system atrophy predicts future dynamics of disability progression in a real-world multiple sclerosis cohort. *European Journal of Neurology* **28**, 4153–4166 (2021).
[5] Ibrahim, A. A. *et al.* Inertial sensor-based gait parameters reflect patient-reported fatigue in multiple sclerosis. *Journal of neuroengineering and rehabilitation* **17**, 1–9 (2020).
[6] Manouchehrinia, A. *et al.* Predicting risk of secondary progression in multiple sclerosis: a nomogram. *Multiple Sclerosis Journal* **25**, 1102–1112 (2019).
[7] Pisani, A. I. *et al.* A novel prognostic score to assess the risk of progression in relapsing- remitting multiple sclerosis patients. *European journal of neurology* **28**, 2503–2512 (2021).
[8] Chalkou, K. *et al.* A two-stage prediction model for heterogeneous effects of treatments. *Statistics in medicine* **40**, 4362–4375 (2021).
[9] Ahuja, Y. *et al.* Leveraging electronic health records data to predict multiple sclerosis disease activity. *Annals of clinical and translational neurology* **8**, 800–810 (2021).
[10] Kurtzke, J. F. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss). *Neurology* **33**, 1444–1444 (1983).
[11] Zhao, Y. *et al.* Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PloS one* **12**, e0174866 (2017).
[12] Zhao, Y. *et al.* Ensemble learning predicts multiple sclerosis disease course in the summit study. *NPJ digital medicine* **3**, 1–8 (2020).
[13] Tommasin, S. *et al.* Machine learning classifier to identify clinical and radiological features relevant to disability progression in multiple sclerosis. *Journal of neurology* **268**, 4834–4845 (2021).
[14] Yperman, J. *et al.* Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis. *BMC neurology* **20**, 1–15 (2020).
[15] Rudick, R. A., LaRocca, N., Hudson, L. D. & Msoac. Multiple sclerosis outcome assessments consortium: genesis and initial project plan. *Multiple Sclerosis Journal* **20**, 12–17 (2014).
[16] Multiple Sclerosis Outcome Assessments Consortium. Msoac - critical path institute. URL: https://c-path.org/programs/msoac/ URL visited on 16th January 2023.
[17] Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* **45**, 1–67 (2011).