MSJ Experimental Translational Clinical

Original Research Paper

Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression

Multiple Sclerosis Journal—Experimental, Translational and Clinical

October–December 2019, 1–14

DOI: 10.1177/2055217319885983

Marco TK Law, Anthony L Traboulsee, David KB Li, Robert L Carruthers, Mark S Freedman, Shanon H Kolind and Roger Tam

Abstract

Background: Enhanced prediction of progression in secondary progressive multiple sclerosis (SPMS) could improve clinical trial design. Machine learning (ML) algorithms are methods for training predictive models with minimal human intervention.

Objective: To evaluate individual and ensemble model performance built using decision tree (DT)-based algorithms compared to logistic regression (LR) and support vector machines (SVMs) for predicting SPMS disability progression.

Methods: SPMS participants ($n = 485$) enrolled in a 2-year placebo-controlled (negative) trial assessing the efficacy of MBP8298 were classified as progressors if a 6-month sustained increase in Expanded Disability Status Scale (EDSS) ($\geq$1.0 or $\geq$0.5 for a baseline of $\leq$5.5 or $\geq$6.0 respectively) was observed. Variables included EDSS, Multiple Sclerosis Functional Composite component scores, T2 lesion volume, brain parenchymal fraction, disease duration, age, and sex. Area under the receiver operating characteristic curve (AUC) was the primary outcome for model evaluation.

Results: Three DT-based models had greater AUCs (61.8%, 60.7%, and 60.2%) than independent and ensemble SVM (52.4% and 51.0%) and LR (49.5% and 51.1%).

Conclusion: SPMS disability progression was best predicted by non-parametric ML. If confirmed, ML could select those with highest progression risk for inclusion in SPMS trial cohorts and reduce the number of low-risk individuals exposed to experimental therapies.

Keywords: Artificial intelligence, machine learning, prognosis, secondary progressive multiple sclerosis, decision support techniques, disease progression

Date received: 17 May 2019; Revised received: 23 August 2019; accepted: 9 October 2019

Introduction

The ability to accurately predict disability progression may lead to an improved understanding of multiple sclerosis (MS) pathogenesis, facilitate faster treatment development, and inform both patient and physician treatment decisions. Selecting individuals predicted to be at high risk of progression within the near future for clinical trials may allow for shorter trial durations as well as reduce the number of individuals exposed to experimental therapies. This is particularly important in secondary progressive multiple sclerosis (SPMS), where disability progression is independent of relapses and treatment options are limited.[1]

Machine learning (ML) algorithms are data science approaches to building predictive models that are able to learn patterns and relationships within data while requiring minimal human intervention. In MS, the application of ML thus far has mainly been for classifying participants into the different disease stages (e.g. clinically isolated syndrome (CIS), relapsing–remitting multiple sclerosis (RRMS), and SPMS),[2–4] or for predicting transition from CIS to clinically definite MS,[5–7] and less for predicting disability progression. One study

Correspondence to:
Marco TK Law,
MS/MRI Research Group, Djavad Mowafaghian Centre for Brain Health, Room 450-B, 3rd Floor, 2215 Wesbrook Mall, Vancouver BC, V6T 2B5 Canada.
marcolaw@msmri.medicine.ubc.ca

Marco TK Law, School of Biomedical Engineering, The University of British Columbia, Vancouver, BC, Canada

Anthony L Traboulsee, Department of Neurology, The University of British

Columbia, Vancouver, BC, Canada

David KB Li, Department of Radiology, The University of British Columbia, Vancouver, BC, Canada

Robert L Carruthers, Department of Neurology, The University of British Columbia, Vancouver, BC, Canada

Mark S Freedman, Department of Medicine, University of Ottawa and The Ottawa Hospital Research Institute, Ottawa, ON, Canada

Shanon H Kolind, Department of Radiology, The University of British Columbia, Vancouver, BC, Canada

Roger Tam, School of Biomedical Engineering, The University of British Columbia, Vancouver, BC, Canada

showed that an ensemble of 10 support vector machines (SVMs) outperformed logistic regression (LR) for predicting disability progression (defined by an Expanded Disability Status Scale (EDSS) increase of 1.0) within 5 years in individuals with EDSS <4.[8] SVMs map the original data to a higher dimension so that it is more linearly separable by a decision plane; linear SVMs (LSVMs) used in the aforementioned study maps data to a higher dimension using a linear transformation. Unlike LR, which fits a linear model to all data points, the decision plane of SVMs is defined by a subset of the data and does not require distributional assumptions.[9]

A benefit of ML is that it can more flexibly model nonlinear relationships. Whereas parametric models like LR and LSVMs place assumptions on the characteristics of the input data, non-parametric models such as the decision tree (DT) do not. Starting from a labeled set of data (parent node), decision rules are learned to split the data into groups (child nodes) that are each "purer" in class composition than the parent node.[10] Each child node then becomes a parent node and the process is repeated until stopping criteria are met. To classify new data using DTs, the learned decision rules are applied.

Ensemble models combine the predictions of multiple models to produce a weighted prediction, similar to humans seeking multiple opinions before making a decision.[11] As a result, ensemble models are less prone to overfitting and generalize better to new data. The random forest (RF) is an ensemble of DTs trained on randomly selected subsets of features from the original dataset.[12] AdaBoost-DT (AdB) is a DT-based ensemble model trained using the AdaBoost algorithm which sequentially trains a set of weak models with class weights determined by misclassifications of the preceding model.[13] The purpose of this study was to evaluate the predictive performance of individual (DT) and ensemble non-parametric (RF and AdB) models trained using the DT algorithm, compared to the individual and ensemble models trained using LR and LSVM algorithms, for prediction of EDSS progression in SPMS on data withheld from model training (i.e. generalizability) and to establish a starting point for predicting SPMS progression using several ML methods.

Materials and methods

Study population

A 2-year randomized, double-blind, placebo-controlled phase III study with participation from 47 centers across 10 countries evaluated the efficacy and safety of MBP8298 in participants diagnosed with SPMS.[14] Of the 612 randomized participants, 539 (88%) completed the study. MBP8298 did not provide a clinical benefit when compared to placebo.

EDSS score was collected every 3 months for 24 months to identify progression, and baseline Multiple Sclerosis Functional Composite (MSFC) component scores – the 9-Hole Peg Test (9HP), Timed 25-Foot Walk (T25W), and Paced Auditory Serial Addition Test (PASAT) – were used. The MSFC component Z-scores were standardized to the Task Force Dataset.[15] T2 lesion volume (T2LV) and brain parenchymal fraction (BPF) were extracted by blinded radiologists and technologists from magnetic resonance imaging (MRI) studies.

Data from both control and treatment arms of the MBP8298 study was filtered to remove participants with multiple missing visits or data entries at any given visit. These include participants that did not have a complete set of baseline clinical scores (EDSS, MSFC, 9HP, T25W, PASAT) or missing baseline T2LV or BPF. Imputation was not performed for participants missing multiple data entries for several reasons. Imputation would require assumptions on the underlying population distribution. Additionally, within a short time-frame, longitudinal clinical and MRI measurements are noisy. Therefore, imputing missing temporal values is unlikely to accurately approximate the true value.

Participants were categorized as either having confirmed disability progression (CDP+) or not (CDP−). Individuals were classified as having confirmed disability progression only if and only if a 6-month sustained-increase in EDSS ($\geq$1.0 or $\geq$0.5 for baseline EDSS $\leq$5.5 or $\geq$6.0 respectively) was observed. As the study concluded at 24 months, participants with an EDSS increases between 18 months and 24 months could not be verified at 6 months for sustained increase and were classified as non-progressors.

Study design

Predictors of progression. Baseline clinical predictors included T25W, 9HP, and PASAT, standardized to the Task Force Dataset,[15] and EDSS. Demographic variables included disease duration (time since first MS diagnosis), age, and sex. Baseline MRI variables included T2LV (mm^3) and

BPF. Longitudinal data was available but the time points overlapped with our prediction target window and was therefore not included.

Tenfold cross-validation. Generalizability was estimated using tenfold stratified cross-validation (10-CV) to train and evaluate the performance of each model. For each 10-CV, the data was split into 10 non-overlapping subsets that had approximately the same prevalence of progression as the original sample; this allowed for 10 cycles of training and validation. In each cycle, nine subsets (90%) were used for training the model (training data) while the remaining subset (10%) was used to assess model performance (validation data).

Data processing. Individual features in each 10-CV cycle's training data were scaled using an approach robust to outliers by first removing the feature's median, then scaling the feature by its interquartile range. Input features in the validation data of each 10-CV cycle were then transformed in the same manner using the statistics of the training data. Scaling of the data was necessary to allow for comparison of predictor importance in LR and SVM using their model coefficients which would otherwise be affected by differing feature magnitudes.

Class imbalance. The dataset has more CDP− than CDP+. To prevent models from preferentially predicting non-progression, random under-sampling was applied to the training data in each 10-CV cycle to balance class representation. Random under-sampling randomly selects CDP- participants to exclude from training so that data presented to the model has equal class representation. Random under-sampling was not applied to the validation data – this allowed for models to be evaluated on datasets that reflected the prevalence of progression in the study population. Independent models were trained on one randomly under-sampled training set, while individual classifiers of each ensemble model were trained on a different, randomly under-sampled training set.

Models for predicting disability progression. We evaluated the performance of independent models trained using two parametric algorithms, LR and linear kernel SVM, and one non-parametric algorithm, the DT. Additionally, we evaluated ensemble models constructed with the aforementioned algorithms: an ensemble LR (ensLR), ensemble LSVM (ensLSVM), RF, and AdB. All models were trained and validated using Scikit-learn 0.20.2 in Python 3.6.[16]

Hyperparameters used for training the individual models were chosen using a fivefold nested cross validation; a fivefold cross-validation grid search within each training dataset of the 10-CV cycles identified ten ideal sets of generalizable hyperparameters that minimized overfitting. Bootstrapping ($n = 2000$) was then applied on each hyperparameter to select the final value used for model training.

The penalty parameter for the individual LSVM was chosen to be 0.81 from a linear search grid from 0.01 to 1.00 with steps of size 0.01. DT node splitting required each child node to contain a minimum of 5% of the total number of training samples and was chosen from 5%, 10%, or 15%.

Ensemble models were constructed using hyperparameters chosen for the individual models. The number of classifiers in each ensemble was selected from three possible choices (2, 5, or 10) classifiers using the same fivefold nested cross-validation and bootstrapping procedure. The final ensLR was constructed with two LR classifiers, and the ensLSVM was constructed using 10 LSVMs. All three choices for RF (using up to a maximum of eight randomly selected input features) and AdB yielded similar performance and so the simplest models (two-classifier ensembles) were chosen.

Evaluation of model performance

Identifying progressors and non-progressors. The ability of each algorithm's trained model to identify progressors and non-progressors can be assessed using sensitivity (true positive rate) and specificity (true negative rate) metrics that are defined as follows:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

A tradeoff exists between sensitivity and specificity that can be visualized in a model's receiver operating characteristic (ROC) curve. The ROC curve plots sensitivity versus $1 -$ specificity and is useful for determining the optimal threshold for classification.[17] The area under the ROC curve (AUC) is a

better measure of performance than accuracy particularly in class-imbalanced problems,[18] and was used as the primary outcome for algorithm comparison. An AUC of 50% indicates no better than random separation, AUC of 0% indicates inversed class separation (i.e., all CDP+ classified as CDP−, and vice versa), while an AUC of 100% indicates perfectly separated classes.

In order to compare the sensitivity and specificity of the various algorithms, models were first optimized using the ROC convex hull method to identify the thresholds that best balanced the sensitivity-specificity trade-off with respect to the training data.[19] Probabilistic predictions made on the validation data were then converted to binary predictions using the identified thresholds to compute sensitivity and specificity.

Predicting progression and non-progression. To assess predictive performance for both progression and non-progression, predictive values and change in pre- to post-test probabilities were used. Positive predictive value (PPV) and change in pre- to post-positive test predictive value (ΔPPV) are defined as:

$$\mathrm{PPV} = \frac{\mathrm{True\ Positives}}{\mathrm{True\ Positives} + \mathrm{False\ Positives}}$$

$$\Delta\mathrm{PPV} = \mathrm{PPV} - \mathrm{Prevalence}_{\mathrm{CDP+}}$$

PPV describes the probability of progression when an individual is predicted to progress. The ΔPPV shows the change in probability that an individual predicted to progress will progress compared to the baseline likelihood defined by the prevalence of progression.

Model performance in predicting non-progression was evaluated using the negative predictive value (NPV) and change in pre- to post-negative test probabilities (ΔNPV), defined as follows.

$$\mathrm{NPV} = \frac{\mathrm{True\ Negatives}}{\mathrm{True\ Negatives} + \mathrm{False\ Negatives}}$$

$$\Delta\mathrm{NPV} = \mathrm{NPV} - \mathrm{Prevalence}_{\mathrm{CDP-}}$$

NPV is the proportion of predicted non-progressors that did not progress. ΔNPV is the change in probability that an individual predicted to be CDP− does not progress compared to the baseline likelihood of non-progression defined by the prevalence of non-progression.

Predictor contribution to model training

In addition to model performance on predicting progression, we examined whether there were qualitative differences in predictor contributions for each trained model as well as the variance in predictor importance across the cross-validation folds. The contribution C of each predictor x in individual and ensemble LR and LSVM models were calculated from the model coefficients c and represented as a percentage:

$$C(x) = \frac{|c_x|}{\sum_{i=0}^{8}|c_i|} \times 100\%$$

RF and AdB predictor contributions were determined by the impact of each predictor on decreasing the impurity at each node; this was extracted from the model at the end of training.

Statistical analysis

Comparison of AUC was performed using Sun and Wu's fast implementation of DeLong's algorithm for comparing correlated AUCs with generalized U-statistics.[20,21] Sensitivity and specificity were compared using the McNemar χ^2 test.[22] PPV and NPV of each algorithm were compared by their predictive values relative to the other models.[23] Changes in pre- to post-positive and negative test probabilities were compared to positive and negative prevalence using one-sample tests of proportions. A significance threshold of 0.05 was used for all comparisons. All analyses were performed in MathWorks MATLAB R2018a.

Results

Study demographics and predictor characteristics

A total of 54 participants (10%) were removed from our study due to missing data, resulting in a study cohort of 485 SPMS participants. The missing diagnosis duration for one participant was replaced by the mean duration of the study cohort. Of the 485 participants, 415 participants experienced an EDSS increase, but only 115 were CDP+. Overall, 370 (76.3%) were CDP- and 115 (23.7%) were CDP+. The baseline characteristics for the final 485 participants in our study population can be found in Table 1.

Table 1. Baseline predictor characteristics of the study sample.

	CDP+ ($n = 115$)	CDP− ($n = 370$)	Overall ($n = 485$)
Demographical features			
# of females	74 (64.3%)	237 (64.1%)	311 (64.1%)
Mean age [years] (SD)	50.3 (8.2)	51.1 (7.9)	50.9 (8.0)
Mean duration[a] [years] (SD)	9.1 (4.4)	9.3 (5.1)	9.3 (5.0)
Clinical features			
Median EDSS (25th, 75th %tile)	6.0 (4.5, 6.0)	6.0 (4.5, 6.5)	6.0 (4.5, 6.5)
Mean T25W[b] [Z] (SD)	0.08 (1.52)	0.05 (1.54)	0.06 (1.54)
Mean 9HP[b] [Z] (SD)	−0.02 (0.93)	0.07 (0.95)	0.05 (0.95)
Mean PASAT[b] [Z] (SD)	0.05 (1.02)	0.01 (1.00)	0.02 (1.01)
MRI biomarkers			
Median T2LV [mm^3]	10,403.9	9012.0	9321.4
(25th, 75th %tile)	(3392.5, 19796.4)	(3730.3, 19889.3)	(3621.6, 19872.8)
Mean BPF (SD)	0.7559 (0.0473)	0.7520 (0.0474)	0.7530 (0.0476)

[a]Disease duration (time since first MS diagnosis).
[b]Standardized to the Task Force Dataset.[14]
Note. Bold face highlights the statistically significant $p < 0.05$ findings.

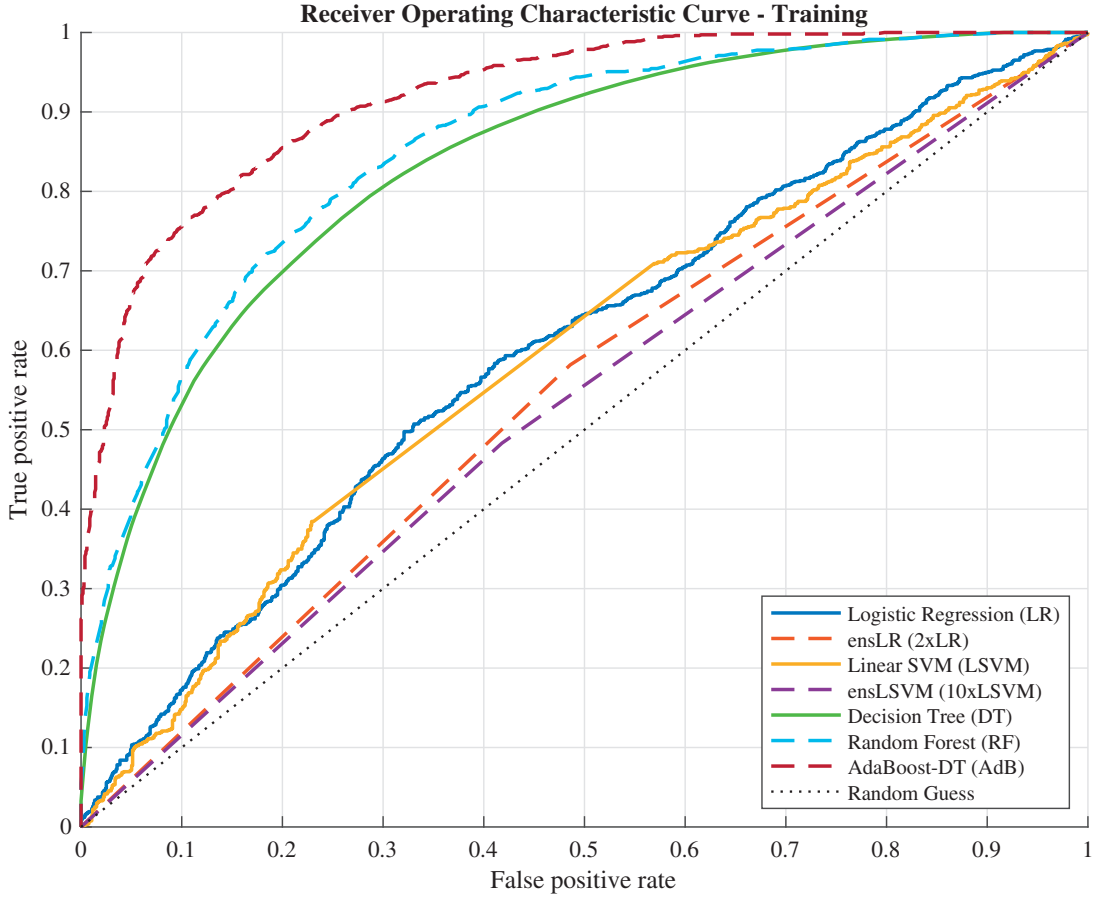

Figure 1. Training receiver operating characteristic curve for individual and ensemble models using logistic regression and linear SVM, and decision tree algorithms

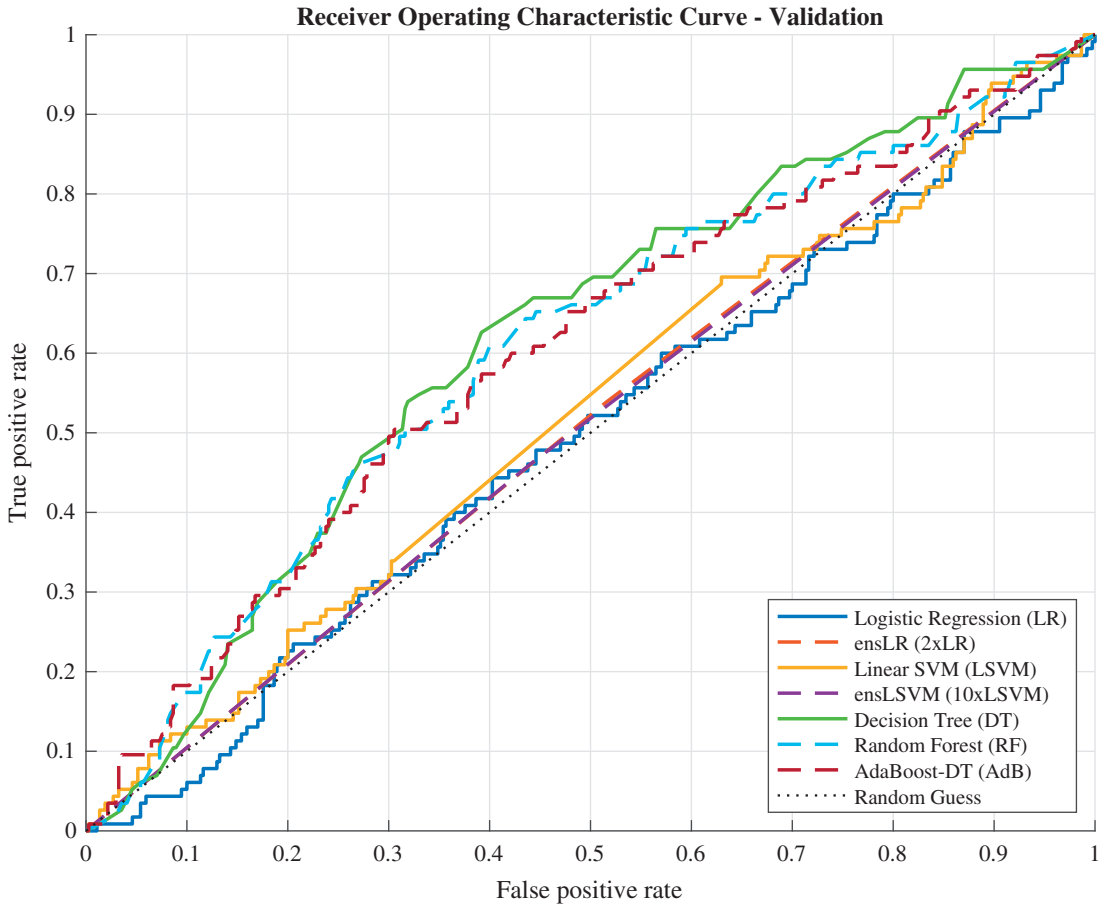

Figure 2. Validation receiver operating characteristic curve for individual and ensemble models using logistic regression and linear SVM, and decision tree algorithms

ROC curves

Parametric models and their ensemble counterparts did not fit the training data as well as the non-parametric models did (Figure 1). This was reflected in model validation ROC curves (Figure 2).

Overall model performance

AUCs summarizing the validation ROC curves in Figure 2 can be seen in Table 2. All non-parametric models outperformed parametric models. No differences were observed between the parametric models or between the non-parametric models.

Optimal classification thresholds were identified to be 49.8%, 50.0%, 49.8%, and 50.0% for LR, ensLR, LSVM, and ensLSVM, and 53.7%, 53.1%, and 52.7% for DT, RF, and AdB. Sensitivity and specificity can be seen in Tables 3 and 4, respectively. Trade-offs between sensitivity and specificity are noticeable in the parametric models, with the model either identifying more CDP+ and less CDP− (as in the ensLR and LSVM) or vice versa (in LR and ensLSVM).

Predicting progression and non-progression

Non-parametric DT models outperformed solo and ensemble LSVM and LR models in PPV. No significant ΔPPV was observed in any parametric models while all DT-based models achieved significant pre- to post-positive test probabilities. These findings are summarized in Table 5.

For NPV, only DT and RF performed significantly better than parametric models. The only parametric model with a significant ΔNPV was LSVM. All the non-parametric models DT, RF and AdB achieved significant ΔNPVs. These findings are summarized in Table 6.

Predictor contribution to model training

Qualitative differences were found in predictor importance between the parametric models and non-parametric DT models. Most notably, T25W

Table 2. Area under the curve (AUC) of individual and ensemble models constructed using logistic regression, SVM, DT algorithms, and comparisons to other models.

Reference model	AUC		% AUC difference[a] (p-value)[b] [95% confidence interval] Comparison Model					
	%	SD	ensLR	LSVM	ensLSVM	DT	RF	AdB
LR	49.5	3.1	1.7 (0.595) [−1.9, 5.3]	2.9 (0.107) [−2.6, 8.4]	1.6 (0.612) [−2.2, 5.3]	**12.3 (0.002) [10.2, 14.4]**	**11.2 (0.006) [9.1, 13.3]**	**10.7 (0.007) [9.3, 13.1]**
ensLR	51.1	2.7		1.2 (0.703) [−2.2, 4.7]	−0.1 (0.965) [−3.4, 3.1]	**10.6 (0.008) [10.6, 10.6]**	**9.5 (0.019) [9.3, 9.8]**	**9.0 (0.0251) [8.2, 9.8]**
LSVM	52.4	3.1			−1.4 (0.653) [−5.1, 2.4]	**9.4 (0.022) [7.8, 11.1]**	**8.3 (0.043) [6.4, 10.2]**	**7.8 (0.058) [5.9, 9.7]**
ensLSVM	51.0	2.7				**10.7 (0.005) [9.2, 12.3]**	**9.7 (0.012) [7.9, 11.5]**	**9.2 (0.015) [7.0, 11.3]**
DT	61.8	3.0					−1.1 (0.460) [−6.6, 4.4]	−1.6 (0.487) [−6.6, 3.4]
RF	60.7	3.1						−0.5 (0.843) [−5.4, 4.3]
AdB	60.2	3.1						

[a]Difference is comparison model AUC minus reference model AUC.
[b]*P*-value obtained using DeLong's algorithm for comparing AUC.[20,21]

contributed very little to the training of parametric models (<3%) while contributing much more to DT models (>15%). Sex contributed more to parametric model training (>9%) than non-parametric models – only contributing 0.8% to DT training, 0.4% to RF training and 1.0% to AdB training. Table 7 summarizes the findings. A plot of the feature contributions to the training of each model is shown in Figure 3 and illustrates the difference in T25W and sex predictor importance on the models examined.

Discussion

In our study population of 485 SPMS participants, we found that DT-based non-parametric models outperformed LR typically seen in data science and linear kernel SVM in separating CDP+ from CDP- (AUC), CDP+ predictive accuracy (PPV), and CDP− predictive accuracy (NPV). In fact, the ROC curves show that both parametric models did not fit the training data well, with LR having identified less than half of progressors and non-progressors.

We observed that there were no significant differences in performance between ensLSVMs, an independent LSVM, and LR. These findings are consistent with those by Zhao et al. when using only baseline features.[8] DT-based models were not restricted to linear relationships and outperformed individual and ensemble LR and LSVMs in predictive accuracies. No statistically significant differences were observed between the non-parametric methods examined. All DT-based models achieved positive pre- to post-test probabilities.

Despite improvements in PPV and NPV demonstrated by DT-based models, significant improvements over parametric models were observed in specificity measures but not sensitivity measures. This may be due to both the sensitivity-specificity trade-off, and

Table 3. Sensitivity performance at optimal classification thresholds of individual and ensemble models constructed using logistic regression, LSVM, DT algorithms, and comparisons to other models.

			% Sensitivity difference[a] (p-value)[b] [95% confidence interval]					
Reference model	Sensitivity		Comparison model					
	%	SD	ensLR	LSVM	ensLSVM	DT	RF	AdB
LR	49.6	4.7	4.3 (0.377) [−5.1, 13.8]	**19.1** **(<0.001)** **[11.9, 26.3]**	−3.5 (0.500) [−13.4, 6.4]	8.7 (0.193) [−4.2, 21.6]	9.6 (0.162) [−3.6, 22.8]	3.5 (0.576) [−8.6, 15.5]
ensLR	53.9	4.6		**14.8** **(0.010)** **[3.9, 25.6]**	−7.8 (0.133) [−17.8, 2.2]	4.3 (0.512) [−8,5, 17.2]	5.2 (0.427) [−7.5, 18.0]	−0.9 (0.896) [−13.7, 12.0]
LSVM	68.7	4.3			**−22.6** **(0.001)** **[−32.3, −12.9]**	−10.4 (0.111) [−23.0, 2.2]	−9.6 (0.141) [−22.1, 3.0]	**−15.7** **(0.014)** **[−27.8, −3.5]**
ensLSVM	46.1	4.6				12.2 (0.053) [0.1, 24.3]	**13.0** **(0.048)** **[0.4, 25.7]**	7.0 (0.262) [−5.0, 18.9]
DT	58.3	4.6					0.9 (0.815) [−6.2, 7.9]	−5.2 (0.265) [−14.2, 3.8]
RF	59.1	4.6						−6.1 (0.200) [−15.2, 3.0]
AdB	53.0	4.7						

[a]Difference is comparison model sensitivity minus reference model sensitivity.
[b]P-value obtained using the McNemar χ^2 test.[22]

relatively small validation sets (approximately 48 samples per validation dataset) generated by 10-CV.

LR continues to be the standard approach in modeling binary disability progression in MS, evaluated based on goodness of fit and not on generalizability. However, our findings suggest that the linear assumption for modeling disability progression in SPMS should be questioned and non-parametric methods should be further explored.

Analyzing predictor contributions to parametric model training, we can see that T25W contributed the least to parametric model training. This leads us to hypothesize that there may be a nonlinear relationship present between T25W and progression which cannot be modeled using linear models, particularly since non-parametric models performed better with greater contributions from T25W. Additionally, we found that sex as a predictor had a near-zero contribution on the better-performing nonlinear models.

In most studies of prognostic factors for disability progression, predictive models use statistical approaches such as linear regression for continuous response prediction or LR for binary response prediction,[24] and Cox regression or Kaplan–Meier analyses for survival analysis.[25] Unfortunately, these analyses do not provide any estimation of their generalizability on samples not used for model fitting. For example, LR was used to evaluate brain atrophy and lesion load as prognostic factors for predicting EDSS score at 10 years.[26] R^2 values were reported for model goodness of fit to the data, but no estimation of how the model would perform on data not used for model fitting was provided. Our study evaluated model performance based on their estimated generalizability by validating models on data withheld from training in each cycle of 10-CV.

Table 4. Specificity performance at optimal classification thresholds of individual and ensemble models constructed using logistic regression, LSVM, DT algorithms, and comparisons to other models.

Reference model	Specificity		% Specificity difference[a] (p-value)[b] [95% confidence interval] Comparison model					
	%	SD	ensLR	LSVM	ensLSVM	DT	RF	AdB
LR	51.1	2.6	−2.7 (0.355) [−8.4, 3.0]	**−14.1 (<0.001) [−18.2, −9.9]**	4.9 (0.081) [−0.57, 10.3]	**11.1 (0.002) [4.0, 18.1]**	**10.0 (0.005) [3.1, 16.9]**	**11.4 (0.002) [4.3, 18.4]**
ensLR	48.4	2.6		**−11.4 (<0.001) [−17.1, −5.6]**	**7.6 (0.011) [1.8, 13.3]**	**13.8 (<0.001) [6.8, 20.7]**	**12.7 (0.001) [5.7, 19.7]**	**14.1 (0.001) [7.3, 20.8]**
LSVM	37.0	2.5			**18.9 (<0.001) [13.3, 24.5]**	**25.1 (<0.001) [18.2, 32.1]**	**24.1 (<0.001) [17.1, 31.0]**	**25.4 (<0.001) [18.3, 32.5]**
ensLSVM	55.9	2.6				6.2 (0.083) [−0.8, 13.2]	5.1 (0.159) [−2.0, 12.2]	6.5 (0.066) [−0.4, 13.4]
DT	62.2	2.5					−1.1 (0.500) [−4.2, 2.0]	0.3 (0.920) [−4.9, 5.5]
RF	61.1	2.5						1.4 (0.621) [−4.0, 6.7]
AdB	62.4	2.5						

[a]Difference is comparison model specificity minus reference model specificity.
[b]*P*-value obtained using the McNemar χ^2 test.[22]

While the models developed from this study provide an improvement in performance over the conventional LR model, LSVMs and prevalence-based baseline performance, additional work is required. Progression defined by an increase in EDSS is weighted towards physical impairment. Using a broader or more comprehensive definition that includes changes in cognition may provide different results. As a ML experiment, our sample of 485 is considered small and demonstrates a difficulty in training ML models – the need for large amounts of data. We hypothesize that in a larger dataset, the improvements in PPV and NPV would be better reflected in model sensitivity and specificity. We used a small set of predictors in this preliminary study. The improvement in performance using nonlinear models may be amplified by the inclusion of additional predictors with nonlinear relationships with. This includes experimenting with automated feature detection from MRIs using an advanced ML method known as deep learning which has been used to predict progression in RRMS by analyzing MRIs.[27] In our study, AdB was constructed using simple DTs; we hypothesize that the use of random trees to construct the AdaBoost ensemble could increase predictive performance.

Our work is one of many steps required to develop a clinically-usable prognostic tool. In its current form, the models developed in this study are not clinically useful for prognosticating an individual's disease course. Despite this, the improvements seen in non-parametric algorithms may aid in streamlining clinical trial recruitment and suggest that non-parametric algorithms may be better suited for evaluating the prognostic value of factors of progression.

In the design of clinical trials and statistical testing, balanced designs are preferred over unbalanced design when possible. Balanced designs result in tests with greater statistical power as they give the maximal information regarding treatment

Table 5. Positive predictive value, relativity to other models, and change in pre- to post-positive test probabilities at optimal classification thresholds of individual and ensemble models constructed using logistic regression, LSVM, DT algorithms, and comparisons to other models.

Reference model	PPV %	PPV SD	Relative PPV[a] (*p*-value)[b] [95% confidence interval] Comparison Model ensLR	LSVM	ensLSVM	DT	RF	AdB	Pre- to post-positive test probability (*p*-value)[c]
LR	23.9	1.9	1.02 (0.780) [0.88, 1.20]	1.06 (0.328) [0.95, 1.18]	1.02 (0.790) [0.86, 1.23]	**1.35 (0.001) [1.13, 1.62]**	**1.34 (0.002) [1.11, 1.61]**	**1.27 (0.011) [1.06, 1.54]**	0.2 (0.899)
ensLR	24.5	2.0		1.03 (0.679) [0.88, 1.21]	1.00 (0.989) [0.84, 1.20]	**1.32 (0.002) [1.11, 1.57]**	**1.31 (0.002) [1.10, 1.55]**	**1.24 (0.023) [1.03, 1.50]**	0.8 (0.696)
LSVM	25.3	2.8			0.97 (0.711) [0.82, 1.14]	**1.28 (0.003) [1.09, 1.50]**	**1.27 (0.003) [1.08, 1.48]**	**1.20 (0.035) [1.01, 1.43]**	1.6 (0.559)
ensLSVM	24.5	1.7				**1.32 (0.003) [1.10, 1.58]**	**1.31 (0.005) [1.08, 1.58]**	**1.24 (0.028) [1.02, 1.51]**	0.8 (0.636)
DT	32.4	2.0					0.99 (0.858) [0.90, 1.09]	0.94 (0.448) [0.81, 1.10]	**8.7 (<0.001)**
RF	32.1	2.1						0.95 (0.521) [0.82, 1.11]	**8.4 (<0.001)**
AdB	30.5	1.6							**6.8 (<0.001)**

[a]relative PPV $= \frac{\text{comparison PPV}}{\text{reference PPV}}$.
[b]*P*-value obtained using Moskowitz and Pepe's algorithm.[23]
[c]*P*-value obtained using one-sample test of proportion of reference model compared to positive prevalence of 23.7%.

Table 6. Negative predictive value, relativity to other models, and change in pre- to post-negative test probabilities at optimal classification thresholds of individual and ensemble models constructed using logistic regression, LSVM, DT algorithms, and comparisons to other models.

Reference model	NPV %	NPV SD	Relative NPV[a] (*p*-value)[b] [95% confidence interval] Comparison model ensLR	LSVM	ensLSVM	DT	RF	AdB	Pre- to post- negative test probability (*p*-value)[c]
LR	76.5	1.9	1.01 (0.758) [0.96, 1.06]	1.03 (0.177) [0.98, 1.09]	1.01 (0.826) [0.96, 1.06]	**1.08 (0.014) [1.02, 1.15]**	**1.08 (0.016) [1.01, 1.15]**	1.06 (0.054) [1.00, 1.12]	0.2 (0.905)
ensLR	77.2	1.8		1.03 (0.472) [0.96, 1.10]	1.00 (0.922) [0.95, 1.05]	**1.07 (0.032) [1.01, 1.14]**	**1.07 (0.030) [1.01, 1.14]**	1.05 (0.126) [0.99, 1.12]	0.9 (0.628)

(continued)

Table 6. Continued

Reference model	NPV %	NPV SD	Relative NPV[a] (p-value)[b] [95% confidence interval] Comparison model ensLR	LSVM	ensLSVM	DT	RF	AdB	Pre- to post- negative test probability (p-value)[c]
LSVM	79.2	1.4			0.97 (0.371) [0.91, 1.03]	1.04 (0.240) [0.97, 1.12]	1.05 (0.234) [0.97, 1.12]	1.02 (0.527) [0.95, 1.10]	**2.9 (0.034)**
ensLSVM	77.0	2.1				**1.08 (0.013) [1.02, 1.14]**	**1.08 (0.017) [1.01, 1.14]**	1.05 (0.068) [1.00, 1.11]	0.7 (0.749)
DT	82.7	1.8					1.00 (0.969) [0.97, 1.03]	0.98 (0.312) [0.94, 1.02]	**6.4 (<0.001)**
RF	82.8	1.7						0.98 (0.311) [0.94, 1.02]	**6.5 (<0.001)**
AdB	81.1	1.9							**4.8 (0.012)**

[a]relative NPV $= \frac{\text{comparison NPV}}{\text{reference NPV}}$
[b]P-value obtained using Moskowitz and Pepe's algorithm.[23]
[c]P-value obtained using one-sample test of proportion of reference model compared to negative prevalence of 76.3%.

Table 7. Contribution of predictors on the training of logistic regression, ensemble SVM, random forest, and AdaBoost models.

Reference model	Mean % feature contribution to algorithm training[a] (SD) Demographic features Age	Sex	Duration	Clinical features EDSS	T25W	9HP	PASAT	MRI features T2LV	BPF
LR	8.7 (5.2)	9.8 (6.4)	4.6 (3.5)	25.4 (5.3)	2.6 (2.9)	17.7 (6.5)	5.8 (5.8)	7.6 (5.3)	17.6 (6.9)
ensLR	9.2 (5.9)	9.6 (9.7)	4.5 (2.7)	28.6 (4.7)	2.2 (1.5)	23.0 (6.1)	8.0 (5.3)	7.1 (4.3)	7.8 (5.1)
LSVM	7.5 (3.7)	10.6 (6.2)	5.4 (4.1)	26.2 (4.6)	2.4 (3.1)	18.1 (5.4)	6.9 (5.6)	6.2 (3.7)	16.6 (6.5)
ensLSVM	7.9 (3.9)	11.5 (21.9)	5.5 (2.8)	17.3 (8.9)	1.4 (0.6)	22.4 (8.5)	5.7 (3.9)	10.0 (9.0)	18.3 (5.9)
DT	10.0 (8.7)	0.8 (2.5)	7.4 (7.7)	24.6 (8.4)	30.2 (8.7)	8.5 (8.6)	3.5 (5.4)	9.9 (5.8)	5.2 (4.4)
RF	10.6 (7.5)	0.4 (1.3)	7.7 (8.3)	23.3 (7.1)	25.7 (9.2)	8.7 (9.1)	5.8 (5.1)	12.1 (6.3)	5.6 (3.3)
AdB	11.8 (4.5)	1.0 (1.8)	8.3 (4.5)	15.0 (5.4)	18.3 (5.7)	14.7 (6.7)	8.5 (6.2)	10.6 (6.0)	11.8 (3.6)

[a]Mean of feature contribution to model training across 10-fold cross validation.

differences.[28] In unbalanced randomized control trials (RCTs), results often favor new treatments when compared to balanced trials.[29] While control/treatment groups can be balanced, unforeseen group imbalances may arise over the duration of the trial. The ideal RCT should consider time-dependent changes (i.e. progression) in the cohort and reduce potential group imbalances. The identification of those most at risk of disability progression during a trial and most likely to benefit from treatment would improve the efficiency of the trial and the power associated with treatment effect findings.

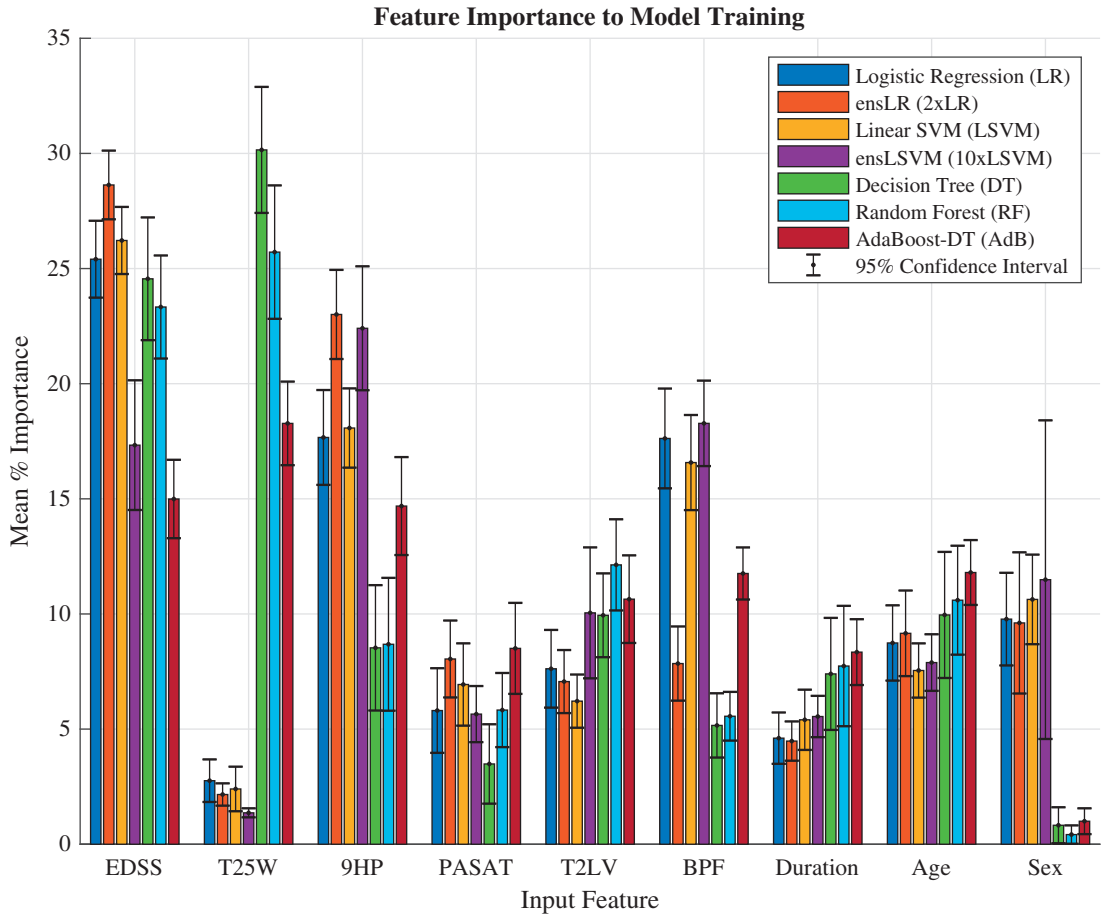

Figure 3. Plot of predictor contribution (with 95% confidence intervals) to independent and ensemble model training using logistic regression, linear SVM, and decision tree algorithms

ML applications in Alzheimer's disease for clinical trial enrichment and design have been shown to enable smaller trials with high statistical power by selecting participants at higher risk of cognitive decline.[30,31] Based on our results, the use of the AdB model would hypothetically reduce the imbalance between progressors and non-progressors by identifying eight more progressors and six fewer non-progressors in every 100 individuals screened for study eligibility. The incorporation of predictive ML models into SPMS clinical trial design may allow those at highest risk of disease worsening to access experimental therapies and yield treatment findings with acceptable statistical power using a smaller study cohort.

Acknowledgements

This investigation was supported (in part) by an endMS Master's Studentship Award from the Multiple Sclerosis Society of Canada.

Conflicts of Interests

The author(s) declared the following potential conflictsofinterest with respect to the research, authorship, and/orpublication of this article: Marco TK Law received research funding support from the Multiple Sclerosis Society of Canada. Anthony L Traboulsee has the following competing financial interests: research funding from Biogen, Chugai, Novartis, Roche, Sanofi Genzyme, and consultancy honoraria from Biogen, Roche, Sanofi Genzyme, Teva Neuroscience. David KB Li has received research funding from the Canadian Institute of Health Research and Multiple Sclerosis Society of Canada. He is the Emeritus Director of the UBC MS/MRI Research Group which has been contracted to perform central analysis of MRI scans for therapeutic trials with Novartis, Perceptives, Roche and Sanofi-Aventis. The UBC MS/MRI Research Group has also received grant support for investigator-initiated independent studies from Genzyme, Merck-Serono, Novartis and Roche. He has acted as a consultant to Vertex Pharmaceuticals and served on the Data and Safety Advisory Board for Opexa Therapeutics and Scientific Advisory Boards for Adelphi Group, Celgene, Novartis and Roche. He has also given lectures which have

been supported by non-restricted education grants from Biogen-Idec, Novartis, Sanofi-Genzyme and Teva. Robert L Carruthers is Site Investigator for studies funded by Roche, Novartis, MedImmune, EMD Serono and receives research support from Teva Innovation Canada, Roche Canada and Vancouver Coastal Health Research Institute. Robert L Carruthers has done consulting work and has received honoraria from Roche, EMD Serono, Sanofi, Biogen, Novartis, and Teva. Mark S Freedman has received a research/educational grant from Genzyme; received honoraria or consultation fees from Actelion, BayerHealthcare, BiogenIdec, Chugai, Clene Nanomedicine, EMD Canada, Genzyme, Merck Serono, Novartis, Hoffman La-Roche, Sanofi-Aventis, Teva Canada Innovation; is member of a company advisory board, board of directors or other similar group of Actelion, BayerHealthcare, BiogenIdec, Hoffman La-Roche, Merck Serono, MedDay, Novartis, Sanofi-Aventis and is on speaker's bureau for Genzyme. Shannon H Kolind has received a research/educational grant funding from Genzyme and Roche. Roger Tam has received research support as part of sponsored clinical studies from Novartis, Roche, and Sanofi Genzyme.

Funding

The author(s) disclosed receipt of the following financial-support for the research, authorship, and/or publication ofthis article: This work was supported by an endMS Personnel Award from the Multiple Sclerosis Society of Canada (grant number 3292).

ORCID iD

Marco TK Law https://orcid.org/0000-0003-1537-8922

References

1. Lorscheider J, Buzzard K, Jokubaitis V, et al. Defining secondary progressive multiple sclerosis. *Brain* 2016; 139: 2395–2405.
2. Zurita M, Montalba C, Labbé T, et al. Characterization of relapsing–remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. *NeuroImage Clin* 2018; 20: 724–730.
3. Ion-Mărgineanu A, Kocevar G, Stamile C, et al. Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features. *Front Neurosci* 2017; 11: 398.
4. Zhong J, Chen DQ, Nantes JC, et al. Combined structural and functional patterns discriminating upper limb motor disability in multiple sclerosis using multivariate approaches. *Brain Imaging Behav* 2017; 11: 754–768.
5. Zhang H, Alberts E, Pongratz V, et al. Predicting conversion from clinically isolated syndrome to multiple sclerosis: an imaging-based machine learning approach. *NeuroImage Clin* 2019; 21: 101593.
6. Bendfeldt K, Taschler B, Gaetano L, et al. MRI-based prediction of conversion from clinically isolated syndrome to clinically definite multiple sclerosis using SVM and lesion geometry. *Brain Imaging Behav* 2019; 13: 1361–1374.
7. Wottschel V, Alexander DC, Kwok PP, et al. Predicting outcome in clinically isolated syndrome using machine learning. *NeuroImage Clin* 2015; 7: 281–287.
8. Zhao Y, Healy BC, Rotstein D, et al. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS One* 2017; 12: e0174866.
9. Cortes C and Vapnik V. Support-vector networks. *Mach Learn* 1995; 20: 273–297.
10. Breiman L, Friedman J, Stone CJ, et al. *Classification and Regression Trees*. 1st ed. Boca Raton, FL: Chapman & Hall/CRC, 1984.
11. Polikar R. Ensemble based systems in decision making. *IEEE Circuits Syst Mag* 2006; 6: 21–45.
12. Breiman L. Random forests. *Mach Learn* 2001; 45; 5–32.
13. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; 55: 119–139.
14. Freedman MS, Bar-Or A, Oger J, et al. A phase III study evaluating the efficacy and safety of MBP8298 in secondary progressive MS. *Neurology* 2011; 77: 1551–1560.
15. Fischer JS, Rudick RA, Cutter GR, et al. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. *Mult Scler J* 1999; 5: 244–250.
16. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; 12: 2825–2830.
17. Florkowski CM. Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin Biochem Rev* 2008; 29(Suppl 1): S83–S87.
18. Ling CX, Huang J and Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. In: *Proceedings of the 18th international joint conference on artificial intelligence (IJCAI'03)*, Acapulco, Mexico, 9-15 August 2003, pp.519–524. Burlington, MA: Morgan Kaufmann Publishers Inc.
19. Bettinger R. *Cost-sensitive classifier selection using the ROC convex hull method*. Cary, NC: SAS Institute, 2003.
20. Sun X and Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett* 2014; 21: 1389–1393.
21. DeLong ER, DeLong DM and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44: 837.

22. Trajman A and Luiz RR. McNemar χ^2 test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scand J Clin Lab Invest* 2008; 68: 77–80.
23. Moskowitz CS and Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clin Trials J Soc Clin Trials* 2006; 3: 272–279.
24. Tripepi G, Jager KJ, Dekker FW, et al. Linear and logistic regression analysis. *Kidney Int* 2008; 73: 806–810.
25. Bewick V, Cheek L and Ball J. Statistics review 12: survival analysis. *Crit Care* 2004; 8: 389–94.
26. Popescu V, Agosta F, Hulst HE, et al. Brain atrophy and lesion load predict long term disability in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2013; 84: 1082–1091.
27. Tousignant A, Lemaître P, Precup D, et al. Prediction of disease progression in multiple sclerosis patients using deep learning analysis of MRI data. In: Cardoso MJ, Feragen A, Glocker B, et al. (eds) *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning*. London: PMLR, 2019, pp.483–492.
28. Berry DA. Sequential statistical methods. In: *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier, PMLR, 2015, pp.634–638.
29. Dibao-Dina C, Caille A and Giraudeau B. Unbalanced rather than balanced randomized controlled trials are more often positive in favor of the new treatment: an exposed and nonexposed study. *J Clin Epidemiol* 2015; 68: 944–949.
30. Ithapu VK, Singh V, Okonkwo OC, et al. Imaging-based enrichment criteria using deep learning algorithms for efficient clinical trials in mild cognitive impairment. *Alzheimer's Dement* 2015; 11: 1489–1499.
31. Ithapu VK, Singh V, Johnson SC. Randomized deep learning methods for clinical trial enrichment and design in Alzheimer's disease. In: *Deep Learning for Medical Image Analysis*. Elsevier, PMLR, 2017, pp.341–378.