Highlights

Explainable time-to-progression predictions in multiple sclerosis

Robbe D'hondt, Klest Dedja, Sofie Aerts, Bart Van Wijmeersch, Tomas Kalincik, Stephen Reddel, Eva Kubala Havrdova, Alessandra Lugaresi, Bianca Weinstock-Guttman, Saloua Mrabet, Patrice Lalive, Allan G Kermode, Serkan Ozakbas, Francesco Patti, Alexandre Prat, Valentina Tomassini, Izanne Roos, Raed Alroughani, Oliver Gerlach, Samia J. Khoury, Vincent van Pesch, Maria Jose Sa, Julie Prevost, Daniele Spitaleri, Pamela McCombe, Claudio Solaro, Anneke van der Walt, Helmut Butzkueven, Guy Laureys, Jose Luis Sanchez-Menoyo, Koen de Gans, Abdullah Al-Asmi, Norma Deri, Tunde Csepany, Talal Al-Harbi, William M Carroll, Csilla Rozsa, Bhim Singhal, Todd A. Hardy, Sudarshini Ramanathan, Liesbet Peeters, Celine Vens, and the MSBase Study Group

- Survival models consider wider disability progression horizons than binary models
- Explainability techniques validate the complex behaviour of black-box models
- Decision support systems can integrate clinician- or engineer-focused explanations

Explainable time-to-progression predictions in multiple sclerosis

Robbe D'hondt[a,b,1,*], Klest Dedja[a,b,1], Sofie Aerts[c,d,e,f], Bart Van Wijmeersch[c,d,e,f], Tomas Kalincik[g,h], Stephen Reddel[i], Eva Kubala Havrdova[j], Alessandra Lugaresi[k,l], Bianca Weinstock-Guttman[m], Saloua Mrabet[n,o], Patrice Lalive[p], Allan G Kermode[q,r], Serkan Ozakbas[s,t], Francesco Patti[u,v], Alexandre Prat[w], Valentina Tomassini[x,y], Izanne Roos[g,h], Raed Alroughani[z], Oliver Gerlach[aa,ab], Samia J. Khoury[ac], Vincent van Pesch[ad,ae], Maria Jose Sa[af,ag], Julie Prevost[ah], Daniele Spitaleri[ai], Pamela McCombe[aj,ak], Claudio Solaro[al,am], Anneke van der Walt[an,ao], Helmut Butzkueven[an,ao], Guy Laureys[ap], Jose Luis Sanchez-Menoyo[aq,ar], Koen de Gans[as], Abdullah Al-Asmi[at,au], Norma Deri[av], Tunde Csepany[aw], Talal Al-Harbi[ax], William M Carroll[q,ay], Csilla Rozsa[az], Bhim Singhal[ba], Todd A. Hardy[i], Sudarshini Ramanathan[bb,bc], Liesbet Peeters[c,d,bd], Celine Vens[a,b], and the MSBase Study Group

[a]*KU Leuven, Dept. Public Health and Primary Care, Kortrijk, Belgium,*
[b]*itec, imec research group at KU Leuven, Kortrijk, Belgium,*
[c]*University MS Centre (UMSC), Hasselt University, Hasselt-Pelt, Belgium,*
[d]*Department of Immunology, Biomedical Research Institute (BIOMED), Hasselt University, Diepenbeek, Belgium,*
[e]*Noorderhart Hospitals, Rehabilitation and MS Centre, Pelt, Belgium,*
[f]*UHasselt, Rehabilitation Research Center (REVAL), Faculty of Rehabilitation Sciences, Diepenbeek, Belgium,*
[g]*Neuroimmunology Centre, Department of Neurology, Royal Melbourne Hospital, Melbourne, Australia,*
[h]*CORe, Department of Medicine, University of Melbourne, Melbourne, Australia,*
[i]*Department of Neurology, Concord Repatriation General Hospital, Sydney, Australia,*
[j]*Department of Neurology and Center of Clinical Neuroscience, First Faculty of Medicine, Charles University in Prague and General University Hospital, Prague, Czech Republic,*
[k]*Dipartimento di Scienze Biomediche e Neuromotorie, Università di Bologna, Bologna, Italia,*
[l]*IRCCS Istituto delle Scienze Neurologiche di Bologna, Bologna, Italia,*
[m]*Department of Neurology, Jacobs MS center for treatment and research, United States,*
[n]*Department of Neurology, LR 18SP03, Clinical Investigation Centre Neurosciences and*

[*]Corresponding author.
Email address: robbe.dhondt@kuleuven.be (Robbe D'hondt)
[1]These authors have contributed equally to this manuscript.

Mental Health, Razi University Hospital, Tunis, Tunisia,
[o] *Faculty of Medicine of Tunis, University of Tunis El Manar, Tunis, Tunisia,*
[p] *Department of Clinical Neurosciences, Division of Neurology, Unit of Neuroimmunology, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland,*
[q] *Perron Institute for Neurological and Translational Science, The University of Western Australia, Perth, Australia,*
[r] *Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Perth, Australia,*
[s] *Izmir University of Economics, Medical Point Hospital, Izmir, Turkey,*
[t] *Multiple Sclerosis Research Association, Izmir, Turkey,*
[u] *Department of Medical and Surgical Sciences and Advanced Technologies, GF Ingrassia, Catania, Italy,*
[v] *Multiple Sclerosis Unit, AOU Policlinico "G Rodolico-San Marco", University of Catania,*
[w] *CHUM MS Center and Universite de Montreal, Montreal, Canada,*
[x] *Institute for Advanced Biomedical Technologies (ITAB), Dept Neurosciences, Imaging and Clinical Sciences, University G. d'Annunzio of Chieti-Pescara, Chieti, Italy,*
[y] *MS Centre, Clinical Neurology, SS Annunziata University Hospital, Chieti, Italy,*
[z] *Division of Neurology, Department of Medicine, Amiri Hospital, Sharq, Kuwait,*
[aa] *Academic MS Center Zuyd, Department of Neurology, Zuyderland Medical Center, Sittard-Geleen, Netherlands,*
[ab] *School for Mental Health and Neuroscience, Department of Neurology, Maastricht University Medical Center, Maastricht 6131 BK, Netherlands.,*
[ac] *Nehme and Therese Tohme Multiple Sclerosis Center, American University of Beirut Medical Center, Beirut, Lebanon,*
[ad] *Department of Neurology, Cliniques Universitaires Saint-Luc, Brussels, Belgium,*
[ae] *Université Catholique de Louvain,*
[af] *Department of Neurology, Centro Hospitalar Universitario de Sao Joao, Porto, Portugal,*
[ag] *FP-I3ID, Instituto de Investigação, Inovação e Desenvolvimento Fernando Pessoa; FCS-UFP, Faculdade de Ciências da Saúde; RISE-UFP, rede de Investigação em Saúde, Universidade Fernando Pessoa, Porto, Portugal,*
[ah] *CSSS Saint-Jérôme, Saint-Jerome, Canada,*
[ai] *Azienda Ospedaliera di Rilievo Nazionale San Giuseppe Moscati Avellino, Avellino, Italy,*
[aj] *Department of Neurology, Royal Brisbane and Women's Hospital, Brisbane, Australia,*
[ak] *University of Queensland, Australia,*
[al] *Department of Neurology, Galliera Hospital, Genova, Italy,*
[am] *Department of Rehabilitation, ML Novarese Hospital Moncrivello, Moncrivello, Italy,*
[an] *Department of Neurology, The Alfred Hospital, Melbourne, Australia,*
[ao] *Department of Neuroscience, School of Translational Medicine, Monash University, Melbourne, Australia,*
[ap] *Department of Neurology, Universitary Hospital Ghent, Ghent, Belgium,*
[aq] *Department of Neurology, Galdakao-Usansolo University Hospital, Osakidetza-Basque Health Service, Galdakao, Spain,*
[ar] *Biocruces-Bizkaia Health Research Institute,*
[as] *Groene Hart Ziekenhuis, Gouda, Netherlands,*

[at]Sultan Qaboos University, Al-Khodh, Oman,
[au]College of Medicine & Health Sciences and Sultan Qaboos University Hospital,
[av]Neurology department, Hospital Fernandez, Capital Federal, Argentina,
[aw]Department of Neurology, Faculty of Medicine, University of Debrecen, Debrecen, Hungary,
[ax]Neurology Department, King Fahad Specialist Hospital-Dammam, Saudi Arabia,
[ay]Sir Charles Gairdner Hospital, Perth, Australia,
[az]Jahn Ferenc Teaching Hospital, Budapest, Hungary,
[ba]Bombay Hospital Institute of Medical Sciences, Mumbai, India,
[bb]Translational Neuroimmunology Group, Kids Neuroscience Centre and Brain and Mind Centre, Faculty of Medicine and Health, University of Sydney, Sydney, Australia,
[bc]Department of Neurology, Concord Clinical School, Concord Hospital, Sydney, Australia,
[bd]I-Biostat, Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium,

---

Abstract

Background: Prognostic machine learning research in multiple sclerosis has been mainly focusing on black-box models predicting whether a patients' disability will progress in a fixed number of years. However, as this is a binary yes/no question, it cannot take individual disease severity into account. Therefore, in this work we propose to model the time to disease progression instead. Additionally, we use explainable machine learning techniques to make the model outputs more interpretable.

Methods: A preprocessed subset of 29,201 patients of the international data registry MSBase was used. Disability was assessed in terms of the Expanded Disability Status Scale (EDSS). We predict the time to significant and confirmed disability progression using random survival forests, a machine learning model for survival analysis. Performance is evaluated on a time-dependent area under the receiver operating characteristic and the precision-recall curves. Importantly, predictions are then explained using SHAP and Bellatrex, two explainability toolboxes, and lead to both global (population-wide) as well as local (patient visit-specific) insights.

Results: On the task of predicting progression in 2 years, the random survival forest achieves state-of-the-art performance, comparable to previous work employing a random forest. However, here the random forest has the added advantage of being able to predict progression over a longer time horizon, with AUROC $> 60\%$ for the first 10 years after baseline. Explainability techniques further validated the model by extracting clinically valid insights

from the predictions made by the model. For example, a clear decline in the per-visit probability of progression is observed in more recent years since 2012, likely reflecting globally increasing use of more effective MS therapies.

Conclusion: The binary classification models found in the literature can be extended to a time-to-event setting without loss of performance, thus allowing a more comprehensive prediction of patient prognosis. Furthermore, explainability techniques proved to be key to reach a better understanding of the model and increase validation of its behaviour.

Keywords: explainable artificial intelligence, survival analysis, multiple sclerosis, disability progression, longitudinal data

1. Introduction

Multiple Sclerosis (MS) is an inflammatory and neurodegenerative autoimmune disease affecting the central nervous system [1]. More specifically, MS affects neural communication, leading to a wide range of symptoms and complaints. Amongst other problems, MS can result in a build-up of severe disability over the years. Disability is often quantified in terms of the *Expanded Disability Status Scale* (EDSS), an MS-specific disability scale introduced by Kurtzke [2]. EDSS scores are composite scores, as they aggregate the impairment to several functional systems (pyramidal, cerebellar, brainstem, sensory, sphincteric, visual, cerebral, and ambulatory). Although EDSS scores have been critiqued because of their overemphasis on walking ability, limited cognitive assessment, and issues with inter-rater reliability [3], they remain the most common measure of (physical) disability and a golden standard for assessing drug effectiveness in clinical trials.

Due to the heterogeneity of the disease course of MS patients, formulating a prognosis is a difficult challenge for clinicians. This has sparked interest in using existing observational and clinical trial data to model disease progression in a data-driven way with statistical and machine learning models. In particular, as evidenced by its role in several recent review papers [4, 5, 6, 7, 8], predicting future EDSS scores is currently a hot topic. However, we note two major limitations in the current line of research, which we discuss in the following two paragraphs.

Currently, predicting future EDSS scores is often done in a rather static way, focusing on predicting progression in a fixed number of years from a baseline visit [4, 5, 6, 7, 8]. This gives only a limited cross-sectional view

of the future progression of a patient, and the binary endpoint makes it impossible to take individual patient severity into account. Only a handful of studies perform a longitudinal prediction instead [9, 10, 11, 12, 13]. They are discussed further in Section 2. Despite these studies, the prediction of *time to progression* remains under-explored in the literature, especially using machine learning models. Amongst other benefits, these models allow us to draw individualised *survival curves* at each patient visit, which give an indication of the speed at which a patient will be progressing in the future. As clinicians are already familiar with survival curves due to their widespread usage for epidemiological purposes, this introduces little to no learning overhead for them.

Additionally, the vast majority of the current machine learning literature for predicting progression in multiple sclerosis is purely concerned with predictive accuracy, ignoring the transparency required of decision support systems in a medical context [6]. Therefore, we also spend considerable effort in this work to explain the predictions of our predictive models to our targeted end users (i.e., clinicians). Specifically, both global and local interpretability are of interest, each supporting distinct but complementary aspects of medical decision-making. Global interpretability provides macroscopic insights into the importance of features for model behaviour. Translating this into a medical setting, this level of understanding can guide healthcare policy or treatment guidelines [14]. On the other hand, local interpretability provides detailed insights into patient-specific variables of interest for individual predictions. For conditions like MS, where individual symptoms can vary widely, making this level of granularity extremely valuable. It enables clinicians to develop personalised treatment plans that have the potential to improve patient outcomes significantly.

The contributions of this manuscript are twofold. First, we define a framework for training time-to-progression survival analysis models on the individual patient visit level. We evaluate this framework on MSBase, a large international data registry containing data from over 44,000 MS patients, and compare it to the recently published setup from De Brouwer et al. [15] that predicts progression in two years on this database. This comparison demonstrates that the performance of such a survival analysis model (when restricted to the static task of predicting progression in a fixed number of years from a baseline visit) is comparable to the current state-of-the-art in progression prediction.

Second, we use existing and novel explainability frameworks to give inter-

pretations to the predictions of these models. Global explainability techniques allow to further compare the existing approach (with a binary endpoint) to the time-to-event approach, by contrasting their feature importances. Additionally, these techniques can validate the behaviour of the proposed model by showing that it has found (non-trivial) patterns in the data that are supported by the literature. On the other hand, local explainability can improve the understanding of the black-box model prediction for clinical end users. In combination with individualised survival curves, this should give clinicians ample reference points to support their decision-making processes.

In summary, this manuscript contributes both methodologically to data science and applicatively to our understanding of MS. Methodologically, we introduce a machine learning framework to analyse the time-to-progression question for MS on the individual patient visit level in a large-scale database, and we show how accompanying explainability procedures can be applied and interpreted. By ensuring a fair comparison, this shows how time-to-event modeling can enrich the classical way of considering progression at a fixed endpoint. Applicatively, we affirm that these black-box machine learning models identify valid patterns in MS prognostication. We go one step further by interpreting non-trivial interaction patterns found by the model, and confirming them through domain expertise and results found in the scientific literature.

The remainder of this paper is organised as follows. In Section 2, we go more in detail on the progression prediction challenge, and we consider the few studies that propose a similar approach to this work. Then, in Section 3, we explain the modelling procedure, including data preprocessing and our definition of time-to-progression. In Section 4, we then compare the proposed approach to the current state-of-the art, both in terms of predictive accuracy and global explainability. We then delve deeper into our proposed approach with variable interactions and local interpretability. Finally, this is followed up by a more general discussion in Section 5 and a conclusion in Section 6.

2. Related Work in Time-To-Progression for Multiple Sclerosis

Machine learning is employed in various aspects of the clinical management of MS, from diagnosis to prognosis and treatment assignment. Here, we are interested in prognosis, more specifically in data-driven predictions of progression in terms of EDSS scores. Most studies using machine learning for predicting EDSS scores do so in a *binary* way, i.e. they predict worsening of

EDSS scores at a fixed endpoint. Depending on the study, this fixed endpoint can be in 2 years [16], 5 years [17], or even 12 years [18].

In this study, our focus shifts from binary classification to survival analysis. An existing line of research uses survival analysis techniques to predict the time to a fixed EDSS milestone (e.g., time to EDSS = 3 [19, 13]), making the time dynamic but keeping the EDSS threshold static instead. However, here we want to make both aspects dynamic by predicting the time to the next worsening of EDSS score. A handful of studies have also considered a similar prediction task and are discussed throughout this section.

A major study is the one by Kalincik et al. [9], where individual treatment response was assessed in terms of the time from treatment initiation to the first EDSS worsening. To this end, a Cox proportional hazards model [20] was built for each treatment for which sufficient data can be extracted. In contrast, we are interested in the time to first EDSS worsening from *any* of the patient visits. Thus, we only build one model for the entire dataset, where the current treatment is included as one of the covariates. This greatly increases the amount of usable data and allows us to consider a predictive modelling approach with machine learning models instead. A machine learning approach, although powerful, comes with the additional challenge of explainability for the end users. As already mentioned in Section 1 and emphasised in a review by Seccia et al. [6], this challenge is currently under-addressed.

In other related predictive studies, we note that the time-to-EDSS-progression question is studied on the patient level rather than on the visit level, i.e., the time-to-event is only defined on one baseline visit. For example, Eshaghi et al. [10] have used Cox proportional hazards models to explore associations between baseline brain tissue volume and time to sustained EDSS progression on a moderately sized cohort of 1214 MS patients. On the other hand, many descriptive statistical studies (e.g. [21]) define a time-to-progression label on every visit rather than just at baseline. A study by Kappos et al. has even shown that this definition is more sensitive to EDSS worsening events [22]. However, for current predictive modeling studies, this definition is often impossible due to data limitations, as either the number of visits is too small or the average follow-up per patient is too short. In particular, for many studies, the average follow-up per patient is limited to at most 3 years [10, 11, 12, 13], whereas the dataset used in this study offers over 14 years of follow-up (see Section 3.1). Nonetheless, such label per visit approach vastly increases the amount of usable data and allows us to answer the time-to-progression question at *any* visit rather than only at the baseline

visit.

3. Methodology

In this section, we delineate our methodology and data analysis pipeline, illustrated visually in Figure 1. We begin with a description of the dataset and our preprocessed input samples in Section 3.1 and the definition of our time-to-event target label in Section 3.2. In Section 3.3, we then describe how the machine learning models are trained and tuned, followed up by a discussion of performance evaluation in Section 3.4. Finally, in Section 3.5, we introduce the explainability toolboxes used to interpret the model predictions.

3.1. Data source and preprocessing

In this paper, data from the MSBase registry [23, 8] is used. MSBase is an international collaboration effort gathering data from over 100,000 MS patients coming from 186 clinics in 43 countries across the globe in a single coherent dataset. We use a September 2020 extraction of the database[2], including all patients aged greater than 18 years who were diagnosed with MS and have at least 12 months of follow-up. This selection contains 44,886 patients with a total of 655,559 visits (of which 523,774 where the EDSS score was measured) across 146 clinics situated in 33 different countries.

Previous work [15] has developed an open source data cleaning pipeline for this database, which we reuse in this paper. In summary, in [15] exclusion criteria were defined related to data quality considerations[3], to not transitioning from clinically isolated syndrome (CIS) to clinically definite MS (CDMS) during the whole follow-up[4], and to date of visit (i.e., all visits before 1970 were discarded). This results in a set of 40,827 patients with 497,586 visits where the EDSS was measured, with an average follow-up per patient of 14.4 years (standard deviation of 9.4 years, median follow-up of 12.4 years).

To make these patient trajectories usable in our machine learning context, they are transformed into input-output pairs from which a model can be

[2]The reason for not using a more recent extraction is to be able to optimally compare to the setup in [15] using a binary label definition.

[3]For example, same-day visits with differing EDSS scores were removed, while same-day visits with equal EDSS scores were de-duplicated. For more details, see [15, Appendix D].

[4]Since this study is about prognosis of CDMS patients, including the CIS patients would only make the target population more heterogeneous. For CIS patients, there is uncertainty regarding whether they will transition to CDMS or not.

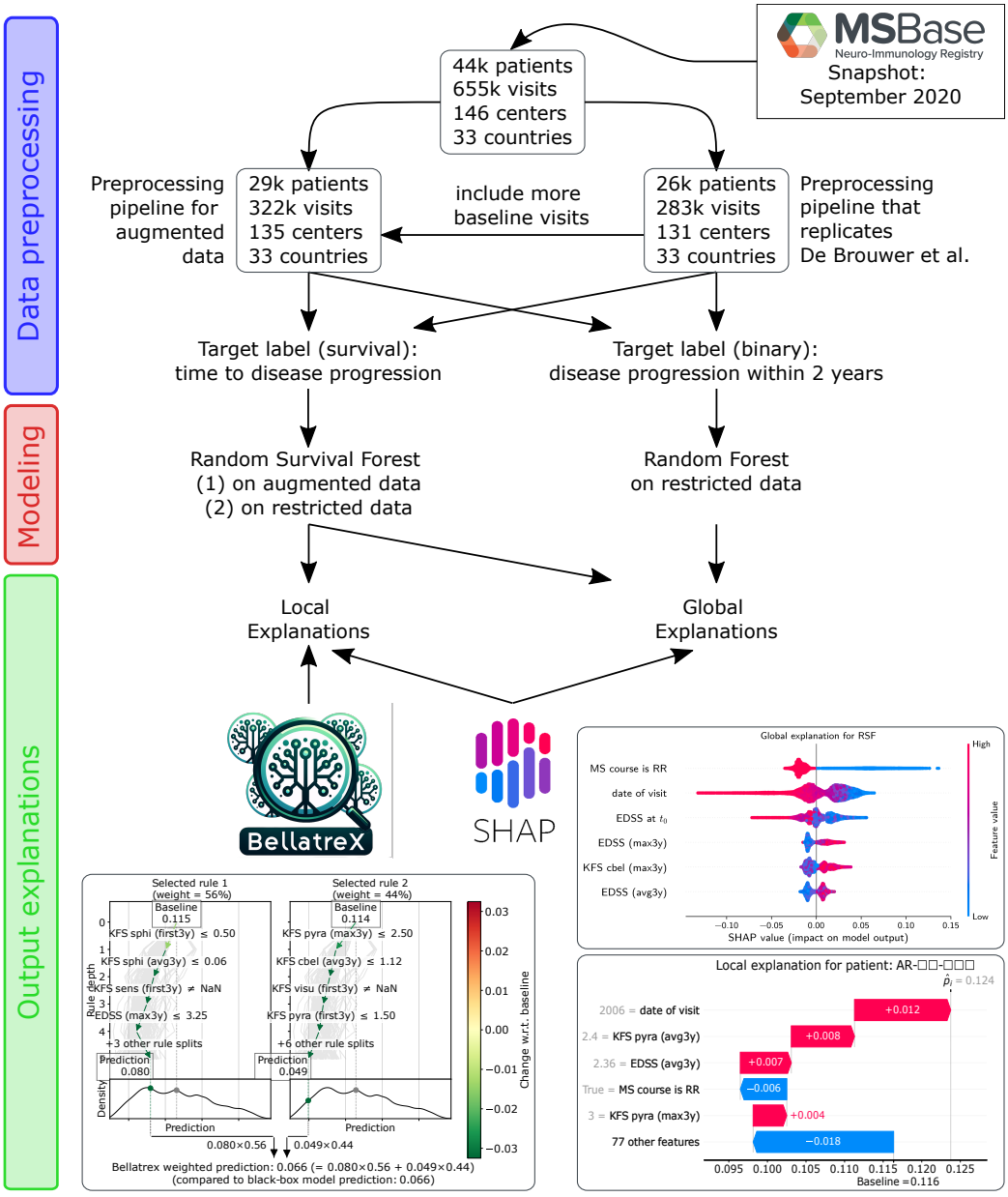


Figure 1: Overview of the pipeline: from raw MSBase data to predictive models with R(S)F, to user-friendly explanations with SHAP and Bellatrex.

trained. These input-output pairs each correspond to a specific patient visit where the EDSS score is measured. We call this visit the *baseline visit*, measured at time t_0 and with EDSS score EDSS_0. However, not all possible visits are valid baseline visits (although all the available visits are still potentially used to define the output label). In De Brouwer et al. [15], where progression in 2 years was studied, only the visits were considered that were after 1990, had an EDSS measurement recorded, had at least 3 EDSS measurements in the 3.25 years before t_0 (to define a clinical history[5]), had at least one EDSS measurement in the 2 years after t_0 (to define the progression label), and in case of progression had at least one EDSS measurement in a relapse-free period[6] at least 2 years after t_0 (to confirm the progression).

In our survival analysis setup, these requirements can be relaxed. Specifically, the last two requirements can be dropped, as they were only necessary to correctly define the binary output label at 2 years after t_0. Consequently, we retain 321,775 baseline visits (i.e., valid t_0), as opposed to the 283,115 visits retained in [15]. For each of these visits, we have 82 features that can be used for prediction: 18 demographic and clinical features (a.o. related to disease onset), 55 EDSS and KFS (Kurtzke Functional Systems) patient history summary statistics (over the past 3 years) and 9 features related to treatment and relapses. A full list can be consulted in Section A. In Table 1, a summary of the cohort characteristics is given.

3.2. Time-to-progression: label definition

The outcome label in this work is the time to the first *significant and confirmed* EDSS progression (also referred to as 'time to progression' for brevity). The EDSS is a numerical scale ranging from 0 to 10 in steps of 0.5. In line with other work in this field [9, 16, 15], the *significance* of progression $\Delta_{0\to1} = \text{EDSS}_1 - \text{EDSS}_0$ from the EDSS score at t_0 to the one at t_1 is defined based on the baseline disability EDSS_0:

$$\Delta_{0\to1} \text{ is significant if } \begin{cases} \Delta_{0\to1} \geq 1.5 \text{ in case} & \text{EDSS}_0 = 0 \\ \Delta_{0\to1} \geq 1.0 \text{ in case} & 0 < \text{EDSS}_0 < 6 \\ \Delta_{0\to1} \geq 0.5 \text{ in case} & \text{EDSS}_0 \geq 6. \end{cases} \tag{1}$$

[5] In [15], the requirement of both 3 and 6 visits is investigated. Here, we only consider the 3 visits case.

[6] Relapse-free is defined as at least 30 days after a relapse.

Feature	Distribution
Dataset size	321775 visits from 29201 patients
MS course	23298 relapsing-remitting (RR) (79.78%) 1953 primary progressive (PP) (6.69%) 2587 secondary progressive (SP) (8.86%) 1363 RR→SP during follow-up (4.67%)
Gender	20654 female (70.73%), 8547 male (29.27%)
Age at onset	32.59 years ± 9.57 years
Age at first visit	41.42 years ± 11.48 years
Follow-up	5.15 years ± 4.76 years
Disease duration at first visit	8.84 years ± 8.14 years
Disease duration at last visit	13.98 years ± 9.07 years
EDSS score	2.97 ± 2.11
EDSS on the visit level	See below

Score	Freq.	Score	Freq.	Score	Freq.
0.0	8.65 %	3.5	5.52 %	7.0	2.14 %
0.5	0.00 %	4.0	8.42 %	7.5	1.24 %
1.0	10.83 %	4.5	3.65 %	8.0	1.07 %
1.5	12.81 %	5.0	2.69 %	8.5	0.37 %
2.0	13.45 %	5.5	3.00 %	9.0	0.15 %
2.5	7.74 %	6.0	6.52 %	9.5	0.02 %
3.0	6.29 %	6.5	5.44 %	10.0	0.00 %

Table 1: Patient characteristics of the preprocessed cohort (on the patient level, unless indicated otherwise). For the continuous variables, mean ± standard deviation over all patients is reported.

If a significant progression event was found, we then also require confirmation of this progression by a third EDSS visit, EDSS_2, at time t_2 in a relapse-free period at most 6 months later (i.e., $\Delta_{0\to 2}$ should also be significant).[7] In case the progression is not confirmed, we move on to the next possible progression event EDSS_1 to define the output label. If no progression events are found, the patient visit is censored at the last available EDSS_1.

In summary, for any patient visit where the EDSS is measured, we define

[7]Technically, progression needs to be confirmed by all visits within a 6 month period after t_1. If there are no visits within the next 6 months, we use the next available visit after t_1 as t_2 instead.

our output as the time until the first significantly increased EDSS measurement (as defined in Equation (1)), with the added requirement that such significant increase is confirmed during the 6 months after the progression event. If the patient never progresses in the available follow-up and hence such time-to-progression cannot be found (or there is no EDSS_2 available to confirm the progression), the output is a time to *censoring* instead, defined as the time until the last EDSS visit of this patient. This indicates that, starting from t_0, the patient does not experience progression before this censoring time, which is still relevant information.

In Figure 2, this time-to-event output label is contrasted against the typical binary label (as found for example in [15]). While the binary label would be defined for a fixed t_1 (e.g., $t_1 = t_0 + 2$ in [15]), the survival analysis label would go over all possible future visits t_1 searching for an actual progression event, and only marks t_0 as a censored visit when such a progression event cannot be found (with the time to censoring then defined as the time to the very last visit). Note in particular in Figure 2 that there are three possible censoring scenarios: (a) none of the visits after t_0 show progression, (b) the only progression event that was found is at the end of follow-up so there are no more visits available to confirm the progression, and (c) all progression events at t_1 are not confirmed by their t_2 visit.

To further exemplify and contrast these definitions of progression, we show an example patient trajectory in Figure 3, with a visualisation of the binary and time-to-event label for each possible baseline visit t_0. Although this is not the most typical patient trajectory[8], the rich follow-up of this patient allows us to showcase many interesting phenomena in the binary and survival analysis label definitions. For example, note the censoring type (b) (compare to Figure 2) observed for all baseline visits for which $t_0 > 18$. Another example: at $t_0 \approx 10.3$ years of disease duration, there is no binary label since there is no t_1 available in the next two years (hence, invalid t_0 for binary modelling). However, a time-to-event *can* be defined, with confirmed progression observed at $t_1 \approx 17.3$ (so time-to-progression equals ≈ 7.0 years). Note that $t_1 \neq 16.1$, since the progression observed at $t_1 \approx 16.1$ is not confirmed by $t_2 \approx 16.5$ as the EDSS has dropped back to 7.5 there. A final example: the second green

[8]The trajectory is atypical as it starts on a relatively high EDSS score of 3.5 early on in the disease duration, and even moves up into the territory of EDSS 8 to 9, with a gap in follow-up of over 5 years.

diamond at $t \approx 4.7$ suggests that there is no progression at $t \approx 6.5$, while the EDSS score is significantly worse there. However, the reason that there is no progression is because the progression is not confirmed: for the visit at $t \approx 7.1$, the EDSS has dropped back to 5.

3.3. Machine learning training setup

To model the time-to-progression label defined in Section 3.2, we employ a Random Survival Forest (RSF) [24]. The RSF is specifically chosen to maximize comparability to the Random Forest from [15], but a comparison to other survival models is given in Section C. As in [15], evaluation of the model is done by 5-fold cross-validation. Each cross-validation split results in a training set ($\approx 80\%$ of the total data), a validation set used for hyperparameter tuning ($\approx 10\%$ of the dataset), and a testing set used for final evaluation ($\approx 10\%$ of the dataset). To closely mimic the conditions of an external validation, this cross-validation procedure is *stratified by clinic*. This means that each training, validation, and testing set cover a separate set of clinics, preventing information leakage at a clinic level. Additionally, we have also implemented a cross-validation procedure *stratified by patient*; this stratification aims to investigate whether the models can achieve a better performance by accessing clinic-specific patterns in the data, while still (as in the stratification by clinic) preventing leakage of visits from the same patient across training, validation, and testing splits.

To align our study with [15] and to facilitate a direct comparison, we adopt the same set of clinics for stratifying the cross-validation in this work. However, since we have relaxed the inclusion criteria for baseline visits (see Section 3.1), our models can be trained on more data (i.e., 321,775 baseline visits in total instead of 283,115). To critically evaluate the impact of this augmented dataset, we conduct an ablation study by also training an RSF on the original dataset (with 283,115 visits) and we refer to this model as the 'restricted RSF'. Both these RSF are then compared to a Random Forest (RF) [25], which is evaluated according to the same procedure.

Tuning of the models is performed with a randomised search with 10 hyperparameter value combinations sampled at random from the grid defined in Section B.2. For RF, the model is chosen that performs the best on the Area Under the Receiver Operating Characteristic (AUROC) [26] when comparing the predictions (probabilities) to the binary progression label as it is defined in [15]. For RSF, the model is chosen that performs the best on the

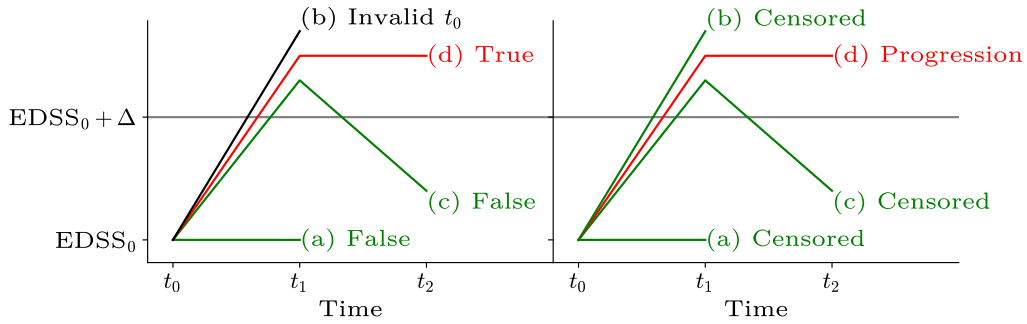


Figure 2: Comparison of the binary (left) and time-to-event (right) label defined on the baseline visit t_0, with significant EDSS progression indicated by Δ. All four possible scenarios with their associated label are shown: (a) no progression at t_1, (b) progression at t_1 but no confirmation visit available, (c) progression at t_1 but not confirmed at t_2, and (d) progression at t_1 and confirmed at t_2.

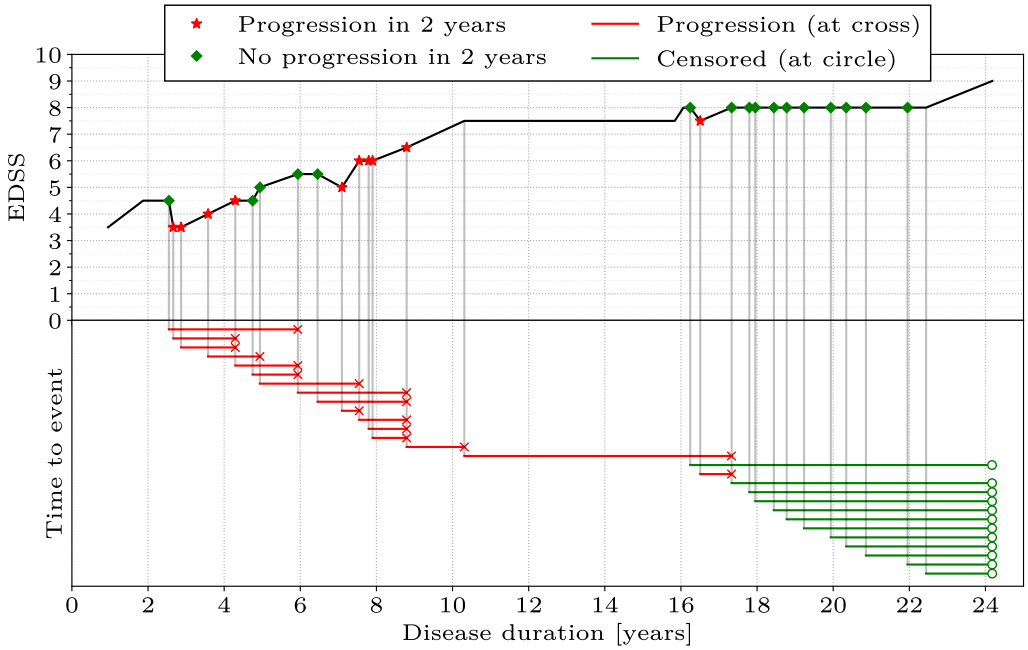


Figure 3: Example patient trajectory with binary progression labels (top) and associated time-to-event labels (bottom) for every possible valid baseline visit t_0. Visits without a label were not considered as valid baseline visits (see Section 3.1 for the validity criteria).

concordance index (C-index) when comparing the predictions (risk scores) to the survival analysis label as it is defined in Section 3.2.

After tuning on the validation set, the model is retrained on the combined train and validation set before the final evaluation on the test set. This final evaluation is done both on the binary and the survival analysis label. A more detailed discussion of the metrics employed for this evaluation is provided in the following section.

3.4. Model evaluation metrics

Here, we define our evaluation strategy to assess predictive performance. In particular, we focus on making a fair comparison between the RF (a model with binary classification outcomes) and the RSF (a model with time-to-event outcomes). A logical first step in this evaluation procedure consists of computing the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall Curve (AUPRC) for the binary label from [15], representing progression in 2 years. For the RF, these metrics are derived from the predicted probabilities. In contrast, the RSF provides a survival curve $\hat{S}_i(t)$ for each observation i, so we first calculate a point estimate probability:

$$\hat{p}_i = 1 - \hat{S}_i(2). \tag{2}$$

This enables a direct comparison of the RSF predictions to the binary outcomes from the RF model.

However, given the longitudinal component of our study, we also take into consideration a time-dependent performance evaluation of the RSF model. Here, in analogy with the evaluation on the binary label, we consider a time-dependent generalisation of the AUROC and AUPRC. This generalisation compares the risk score generated by the RSF to a *binarisation* of the time-to-event label at multiple time points (i.e., all time points in the first 15 years after diagnosis). The binarisation process defines the label to be *true* if the patient progresses before the time point of interest t, *false* if the patient progresses or is censored after t, and *missing* if the patient is censored before t. Note that for $t = 2$, due to this missingness, this binarisation process does not result in the same binary label as used in the evaluation above. For further insights about this difference and its possible effects, refer to Section B.1.

Additionally, we also highlight sensitivity and specificity values for three time points of interest after t_0: $t = 2$ years, $t = 4$ years, and $t = 8$ years.

The specificity (or true negative rate) is defined as the proportion of non-progressing visits that are labelled as progressing for a certain decision threshold. The sensitivity (or true positive rate) is defined as the proportion of progressing visits that are also labelled as progressing for a certain decision threshold. The optimal decision threshold depends on the relative importance of sensitivity versus specificity, and is chosen by the domain expert. In this work, we consider three scenarios: (1) equal importance, (2) sensitivity is twice as important as specificity, and (3) sensitivity is three times as important as specificity. We argue that specificity will never be more important than sensitivity, as the benefit of correctly identifying progressing patients outweighs the benefit of correctly identifying non-progressors.

3.5. Explainability

To explain the predictions of our models, we opted for two post-hoc interpretability toolboxes. The first one, SHAP [27], was chosen due to its widespread adoption and the availability of a fast implementation for RF [28]. The second one, Bellatrex [29], was chosen for its native support for both binary (RF) and time-to-event (RSF) predictions. Note that, to ensure comparability of explanations between RF and RSF, the explanations for RSF are generated for the transformed point estimate probabilistic predictions from Equation (2). In the remainder of this section, we give a short introduction to SHAP and Bellatrex, and we explain the inner workings of these toolboxes in more detail.

SHAP is an interpretability toolbox that uses concepts from game theory to assign a mathematically consistent feature importance to the predicted output of a model. More specifically, given an instance i and the associated model prediction $\hat{y}_i$, SHAP estimates the feature importance vector $\phi_i = \{\phi_{i,0}, \phi_{i,1}, \ldots, \phi_{i,p}\}$, where $\phi_{i,j}$ (for $j \neq 0$) corresponds to the SHAP value for feature j. This value indicates the extent to which feature j contributes, either positively or negatively, to the model's prediction for instance i relative to the baseline[9] prediction $\phi_{i,0}$. For a model prediction $\hat{y}_i$, SHAP thus builds

[9]In this context, 'baseline' refers to the average prediction over all samples (in the training data). Not to be confused with 'baseline visit' used throughout the rest of the paper, which refers to the patient visit at t_0.

a feature importance vector ϕ_i such that

$$\sum_{i=0}^{p} \phi_{i,j} = \hat{y}_i.$$

In Figure 4, we show a schematic representation of the resulting explanation. Note that this approach allows both local interpretability, through individual feature importance vectors ϕ_i, and global interpretability, by aggregating feature importance vectors across instances in the test dataset to understand overall model behaviour.[10]

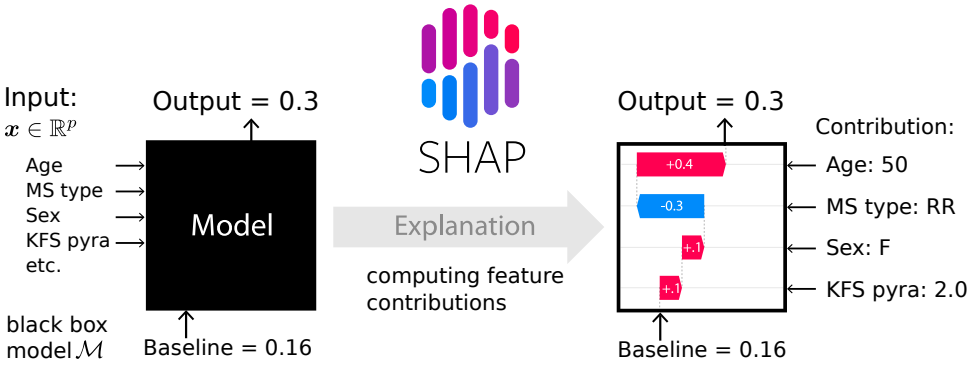


Figure 4: SHAP estimates the contribution of the input features, which cumulatively explain the model's final prediction for a given instance (i.e., local explainability). Figure adapted from https://shap.readthedocs.io/en/latest/.

On the other hand, Bellatrex is an interpretability toolbox specifically designed to explain R(S)F models and offers a perspective on local interpretability complementary to SHAP. As shown in Figure 5, Bellatrex makes use of the internal structure of an R(S)F to select meaningful rules constructed by the model itself and, after a clustering step, select a few rules that are both representative and diverse to show to the end-user as an explanation. Furthermore, the toolbox is designed to ensure that the predictions arising from the extracted rules are closely aligned with the original predictions from the underlying random forest model.[11]

[10] For more details about the method, refer to [27]. For a quick guide on interpreting SHAP explanations, see https://shap.readthedocs.io/en/latest/.

[11] For more details about the method, refer to [29]. For a quick guide on interpreting Bellatrex explanations, see https://itec.kuleuven-kulak.be/a-guide-to-bellatrex/.

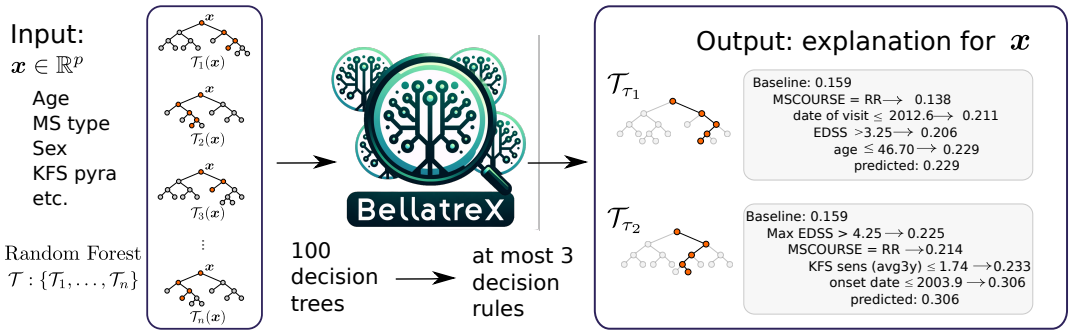


Figure 5: Bellatrex extracts a few accurate and diverse rules from a trained R(S)F and shows them as an explanation.

In this paper, we apply SHAP to obtain global interpretability insights for the RF and RSF model predictions. These global insights shed light on the general patterns found by the model and show which features are overall the most predictive. Next, we employ SHAP and Bellatrex for local interpretability of RSF predictions. These local explanations provide case-specific insights on the prediction process for the instance at hand. We run such local interpretability for two selected patient visits: one with a positive prognosis (no EDSS progression for the next four years) and one with a negative prognosis (confirmed EDSS progression within two years). More details on how these patient visits were selected can be found in Section D.4.

4. Results

In this section, we detail the results of our study, which are divided into two categories: performance-related and explainability-related. Concerning the performance results, we report the binary and time-dependent performance in Section 4.1, particularly focusing on comparing the performance of a Random Forest (RF) predicting whether or not a patient will progress *within 2 years* with a Random Survival Forest (RSF) predicting *when* a patient will progress. As discussed in Section 3.3, these performance results are based on a cross-validation stratified by clinic, to mimic external validation performance. The interested reader can find the results on the patient-based stratification in Section B.3.

As for the explainability results, in Section 4.2, we use SHAP to extract global (i.e., population-wide) interpretability insights about the RF and RSF

models. Additionally, in Section 4.3 we use both SHAP and Bellatrex on the RSF model to illustrate the kind of local explanations that can be derived for specific patient visits. In particular, we investigate one visit with a positive prognosis (i.e., no EDSS progression for the next four years) and one with a negative prognosis (i.e., EDSS progression within two years).

4.1. Performance

The performance results on the binary label from [15] are shown in Table 2. Both in terms of AUROC and AUPRC, the results show that the predictive performance of both RSF models, when limiting the prediction to a 2-year horizon, is similar to the performance of the RF model, with just a 1 percent drop in both metrics (equivalent to 1 standard deviation). In Section D.1, we make this performance analysis also on the subgroup of RRMS patients.

Table 2: Performance results for predicting progression in 2 years. The average and standard deviation over the 5-fold cross-validation are reported. The naive classifier predicts the majority class for all baseline visits.

	AUROC	AUPRC
Naive classifier	0.500 ± 0.000	0.127 ± 0.009
RF	0.706 ± 0.009	0.243 ± 0.021
Restricted RSF	0.695 ± 0.014	0.233 ± 0.021
RSF	0.694 ± 0.009	0.234 ± 0.020

However, by modelling time-to-progression, we consider a more comprehensive longitudinal prediction of patient progression. In Figure 6, we use the Kaplan-Meier estimator to provide a visual representation of the cumulative risk of progression across time in the study population. In particular, this curve shows that any given patient visit has a probability of 50% to not experience progression within the next 8 years. Equivalently, the median time to progression at any baseline visit is 8 years.

Following the binarisation process defined in Section 3.4, a time-dependent AUROC and AUPRC is calculated and shown in Figure 7. From this, we see that overall the RSF with relaxed baseline visit selection criteria (see Section 3.1) outperforms the restricted RSF. Furthermore, for $t = 2$, the RF and RSF again all have similar performance. Note that the RF performance in Table 2 and Figure 7 is different. This is to be expected, given the difference between the binary and binarised label hinted towards in Section 3.4. See Section B.1 for more technical details.

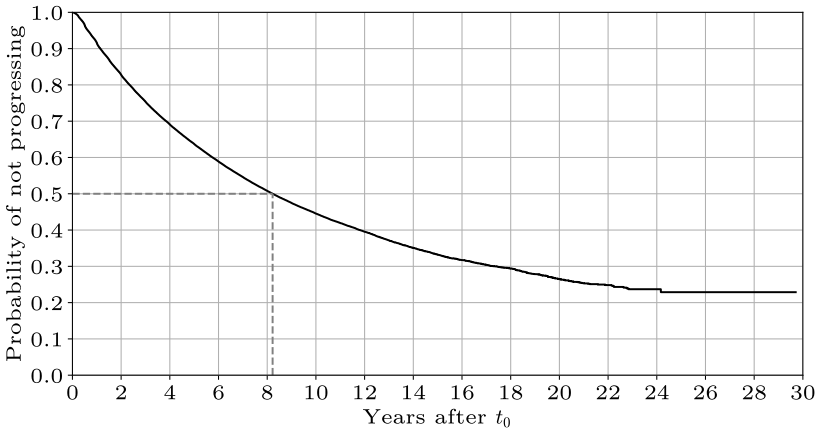


Figure 6: Kaplan-Meier estimator of the population survival curve.

Figure 7 shows a decline in performance over time, which is to be expected. As the prediction time point moves further into the future, predicting time to progression becomes more difficult, and there is also less training data available. In Figure 8, we present a more detailed account of how exactly the performance changes over time based on the ROC curve itself. Interestingly, performance only seems to deteriorate in the upper right corner of the graph. Specifically, upon further inspection, the point of divergence is the classifier threshold probability of 20%. For lower threshold probabilities, the false positive rate (FPR) starts increasing significantly over the years, whereas the true positive rate remains mostly constant (at every decision threshold).

Looking at the binarisation process (see also Section B.1), there are only two possible reasons why the FPR would differ between two time points t_1 and t_2 (assume $t_1 < t_2$): either because of the visits censored in $[t_1, t_2]$ or because of the visits that have an event in $[t_1, t_2]$. Note that both of these groups are ignored in the computation of the FPR at t_2: the first because the label is set to missing, the second because the label is set to true (and FPR only looks at visits labeled as false). Further analysis shows that the first group is larger than the second group (and therefore has more weight), and that the overall FPR for t_1 lies between the (lower) FPR of the first group and the (higher) FPR of the second group. Hence, the visits censored in $[t_1, t_2]$ are the reason for the increased FPR upon their exclusion at t_2: they are easier to classify as 'no progression' at t_1 for the model than other visits. In summary, the decreased performance over time for decision thresholds $< 20\%$

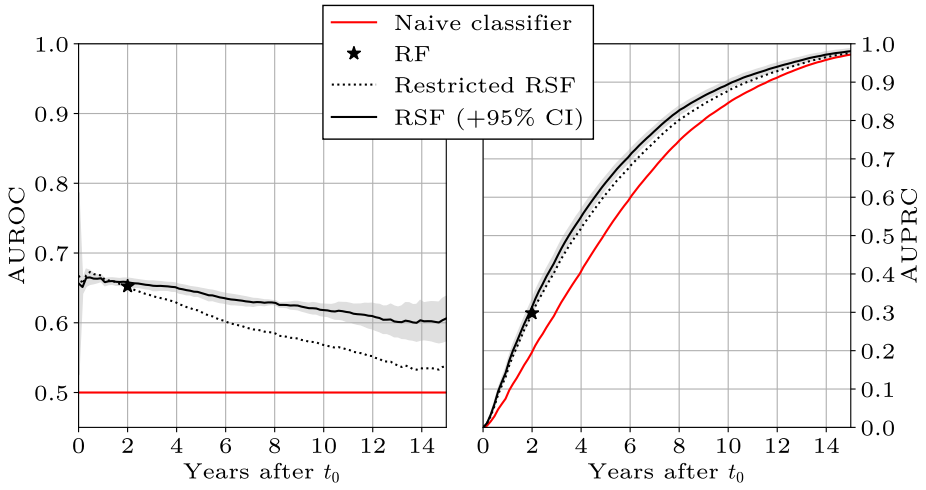


Figure 7: Time-dependent AUROC and AUPRC. The naive classifier predicts at each time point the majority class for all baseline visits. CI = Confidence Interval.

is due to an increase in FPR caused by the exclusion of censored visits as we move our binarisation threshold.

Finally, in Table 3, sensitivity and specificity for 9 different scenarios are reported. Note how the sensitivity increases and specificity decreases as we increase the weighting parameter w. For $w = 1$, the optimal threshold is defined by the intersection of the isocost line $TPR \sim FPR$ with the ROC curve that is closest to the optimal classifier $(FPR = 0, TPR = 1)$. As can be seen from the decision threshold 20% in Figure 8, this part of the ROC curves corresponds to a decreasing TPR and FPR over the years, or a decreasing sensitivity and increasing specificity, in line with the results reported in Table 3. Analogously, for $w = 2$, the optimal threshold is found by intersecting $TPR \sim 2 \cdot FPR$ with the ROC curves. The optimal threshold here is closest to the decision threshold 10% in Figure 8, where sensitivity remains mostly constant over the years but specificity decreases.

4.2. Global explanations

To analyse global (i.e, population-wide) interpretability, we employ the SHAP [27] toolbox. In Section 4.2.1, we perform a univariate analysis for both the RF and RSF models, where we investigate and compare their most predictive features. Then, in Section 4.2.2, we delve deeper into our RSF model with a bivariate analysis, where we investigate how these highly

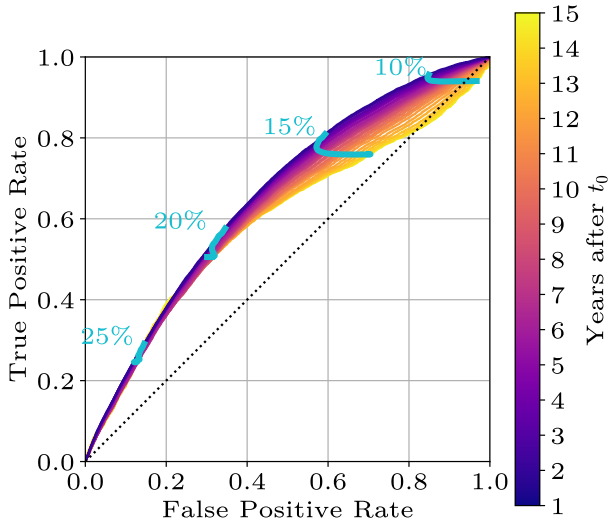


Figure 8: Time-dependent ROC curves for the RSF model, showing in more detail the performance deterioration over time. In cyan, the evolution of TPR/FPR over time for a few decision thresholds is shown.

predictive features interact with other features.

4.2.1. Univariate trends

In Figure 9, the global SHAP explanations for the prediction of each patient visit[12] by the RF and RSF models are shown. Although the direction and magnitude of feature contributions are mostly comparable across the two methods, there are some slight differences. These can be attributed not only to the inherent differences in the predictive tasks (binary versus time-to-event), but also to the difference in inclusion criteria that were used in both cases.

For the binary classification RF model, shown in Figure 9a, the *date of visit*[13] emerges as the most important predictor. More recent visits, denoted by

[12]Technically, only the visits in the test fold of the first train-validation-test split are shown here, but the conclusions remain the same for the test folds in the other splits.

[13]Note that the contribution of 'date of visit' is an aggregate of many factors, such as shifts in the enrolled MSBase population, inclusion of new centres, changes in the use of the EDSS scale, and in the use of treatments. This feature is included here to be able to retrospectively estimate its contribution to the risk of progression. Models aimed at being implemented in current clinical practice (and thus make predictions for new visits) should

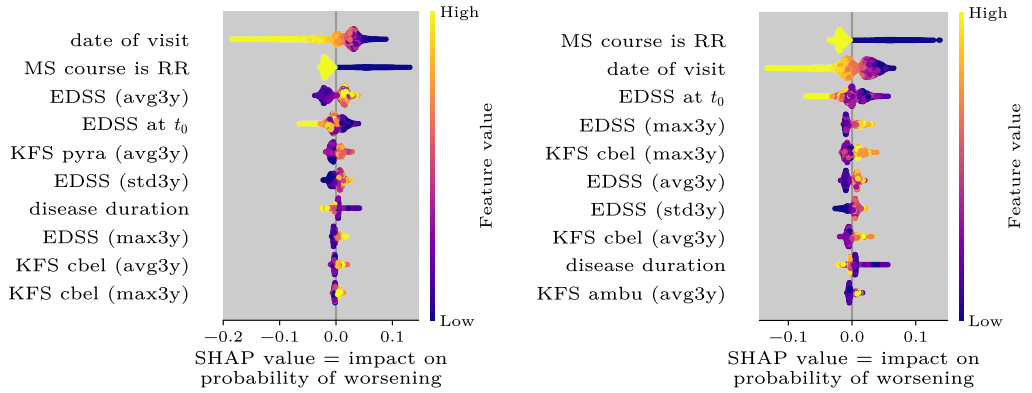
Table 3: Sensitivity and specificity values for 3 time points t after t_0 and three weighting scenarios w based on the relative importance of sensitivity over specificity.

		$t = 2$	$t = 4$	$t = 8$
$w = 1$	Sensitivity	0.683	0.627	0.606
	Specificity	0.555	0.602	0.597
$w = 2$	Sensitivity	0.914	0.902	1.000
	Specificity	0.234	0.229	0.000
$w = 3$	Sensitivity	0.977	1.000	1.000
	Specificity	0.088	0.002	0.000

the yellow dots, have a much lower predicted risk of progression, as evidenced by their negative SHAP values. In particular, recent visit dates have SHAP values down to -0.184, signifying a decrease in predicted risk of worsening of up to 18.4%. This could be related to the increased quality of care of MS patients over the years due to improved code of practice and treatment availability [30]. However, it is important to recognise that the importance of the *date of visit* as a predictor might also be somewhat overestimated due to lack of sufficient follow-up. For the binary progression outcome, this follow-up problem decreases the probability of effectively observing progression in the more recent visits, pushing the RF predictions towards lower predicted risk. In contrast, the time-to-event outcome is less sensitive to the shorter follow-up in more recent visits, as that kind of information is now translated into censored observations. As can be seen in Figure 9b, this results in a reduced predictive impact of this feature, now the second most important predictor and only decreasing the risk up to 13.2%. This may reflect a more nuanced understanding of the disease progression dynamics captured by the time-to-event RSF model.

The most important predictor for the RSF model (and the second most important for the RF model) is related to the current clinical course of the disease for this patient (i.e., the MS subtype). In particular, patients whose MS course is classified as relapsing-remitting MS (RRMS) are associated with a decreased risk of progression with respect to non-RRMS patients, as the yellow dots (i.e., *MS course is RR* is true) indicate negative SHAP contributions. This makes sense, as non-RRMS solely consists of progressive

omit this feature at the training stage.



(a) RF model (binary outcomes)

(b) RSF model (time-to-event outcomes)

Figure 9: Global SHAP explanations, showing the overall impact of each input feature on the predicted probability of progression. For binary features (e.g., *MS course is RR*), *true* corresponds to high feature values, whereas *false* corresponds to low ones. For details about the meaning of the variable names, see the data dictionary in Section A. In particular, (1) KFS scores: pyra = pyramidal, cbel = cerebellar, ambu = ambulation, (2) aggregations: avg = average, std = standard deviation, max = maximum, 3y = over the past three years.

MS in this context, where disability accumulates continuously over time. In particular, this group of progressive MS patients is further subdivided into primary progressive MS (PPMS), which are patients progressive from onset of the disease, and secondary progressive MS (SPMS), which are patients who were initially RRMS but gradually evolved into a progressive phase of the disease. In contrast, RRMS is defined by acute clinical episodes of temporary disease worsening that recovers fully or partly within days or weeks. It is noteworthy, however, that diagnosing a patient as RRMS inherently suggests a non-progression bias. This can be seen as a somewhat circular argument, since a shift to SPMS classification typically follows observed progression, defined by the same clinician who also records this progression.

The third and fourth most predictive features for the RF and RSF models reveal an interesting trend. Whereas a higher average or maximum EDSS score over the past three years is associated with a higher risk of progression, the opposite is true for the current EDSS score at the baseline visit. Although this is seemingly contradictory, it can easily be understood through a *regression to the mean* effect. In a statistical sense, *EDSS at* t_0 is a much more unstable and less representative measurement of the current disability level of a patient than the aggregate score over three years (maximum or average). When

the latter is higher, it signifies that the true underlying disability level of this patient is also high, increasing the risk of regressing back onto this true underlying disability level. This might seem like a rare situation, but several examples of this phenomenon can be found in the example patient trajectory in Figure 3. For example, consider the very first red star visit, which has a high risk of progression (both in the binary and survival analysis setup), and indeed also has a relatively low current EDSS score and a relatively high maximum and average EDSS score over the past three years. Other examples can be found at $t_0 \approx 7.1$ and $t_0 \approx 16.5$.

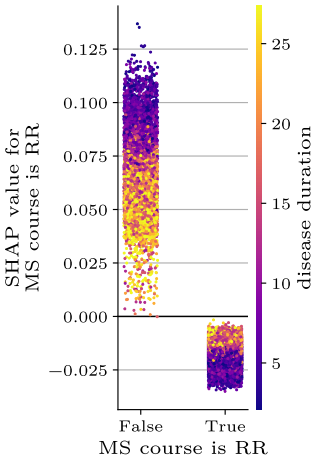
Interestingly, the RF and RSF models seem to disagree for the fifth most important feature. Whereas the RF considers the average *pyramidal* KFS score over the past 3 years more important to predict disability progression within 2 years, the RSF instead considers the maximum *cerebellar* KFS score over the past three years to be more significant in predicting the time to progression. The direction of the effect is the same for both models: higher KFS summary statistic scores have a higher predicted risk of progression. This presents an interesting opportunity for future work to investigate these KFS score contributions in a causal framework. Knowledge of the time-dependent causal contributions of KFS scores towards EDSS progression could guide treatment policy, especially if one of the KFS scores is found out to be more prognostically relevant than the others.

4.2.2. Interaction effects

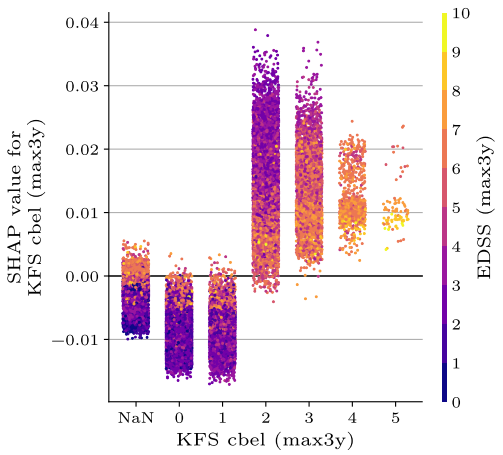
From here, we prioritize the RSF model, given its ability to capture disease progression over time and to handle visits with insufficient follow-up data. To further investigate the univariate trends, we plot the value of each patient visit against its SHAP value for three of the top five most predictive features from Figure 9b: MS course is RR, date of visit, and maximum cerebellar KFS score observed in the past three years before t_0.[14] To highlight variable interaction effects, we then colour code each of these plots based on a second variable, namely the variable that has the highest absolute Pearson correlation $|r|$ with the SHAP values of the plotted variable. The results of this are shown in Figure 10.

The first interaction we analyse, based on the 'MS course is RR' variable,

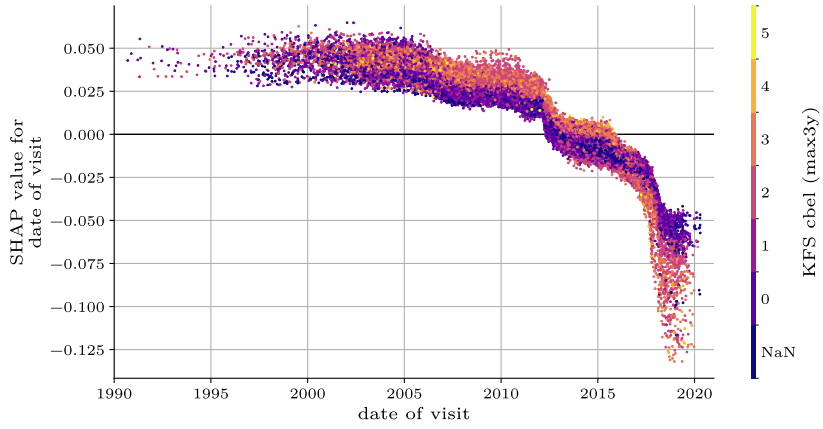
[14]The interactions for the remaining two variables are moved to Section D.3 as they don't bring added value on top of the ones investigated here.



(a) MS course is RR: SHAP values attain $|r| = 0.87$ with disease duration.



(b) KFS cbel (max3y): SHAP values attain $|r| = 0.64$ with EDSS (max3y).



(c) date of visit: SHAP values attain $|r| = 0.60$ with KFS cbel (max3y).

Figure 10: SHAP interaction plots. For each patient visit, a highly predictive feature for the RSF is plotted against its SHAP values. The plots are colour-coded by the variable with the highest absolute correlation $|r|$ with these SHAP values.

is shown in Figure 10a. As already seen in Section 4.2.1 (i.e., Figure 9b), RRMS patients have smaller SHAP values and thus a lower predicted risk of progression than non-RRMS (i.e., progressive here) patients. However, here we are interested in the interaction effect with disease duration, given by the colour coding. We observe that a long disease duration is associated with increased risk of progression for patients with RRMS, whereas remarkably the opposite trend is observed for progressive patients. For RRMS patients, the increasing risk of progression over the disease duration can be explained by the neurological damage from relapses becoming less recoverable as the disease advances [1]. On the other hand, PPMS patients are by their clinical definition progressing from diagnosis, so they experience most progression in the first few years of the disease[15] and are more stable in terms of EDSS scores later in the disease. Additionally, a significant portion of RRMS patients transition to SPMS over time [1], whereby irreversible neurodegeneration gradually becomes more prominent, further explaining this risk increase.[16] This could also partly explain the reversed trend observed for non-RRMS patients, initially consisting mainly of primary progressive MS (PPMS) patients.[17] For longer disease duration, the group of non-RRMS patients is enriched with SPMS patients, who may have a slightly better prognosis than PPMS patients due to the progression already sustained in the RRMS phase.

A similar reversed interaction effect is observed in Figure 10b, where a low maximum EDSS score over the past 3 years is associated with a decreased risk when the maximum cerebellar KFS score is also low (0-1), while it is associated with an increased risk when the maximum cerebellar KFS score is relatively high (2-5). This pattern highlights the direct influence of the cerebellar and pyramidal KFS score on the overall EDSS score, particularly for values of EDSS exceeding 4, where the ambulatory aspect becomes dominant. Furthermore, cerebellar signs early on in the disease is known to be an

[15]Note that disease duration is not the same as the underlying pathology duration.

[16]A patient can still be classified as RRMS even though they have entered the progressive phase of the disease. Before a diagnosis of SPMS can be made, progression must be observed for at least 6 months [31].

[17]The fraction of PPMS patient visits in the non-RRMS group monotonically decreases over almost all years of disease duration, from 90.2% of the visits in the first year of the disease to 52.8% in the eighth year (after which it is no longer the majority fraction), down to 16.8% in the 25th year. Cumulatively speaking: of all non-RRMS visits with disease duration below 13 years (20,227 visits), 50.8% are from PPMS patients.

indicator of poor prognosis [32]. Since a low max EDSS score can be seen as a proxy for low disease duration[18], this further reinforces this finding.

The third and final interaction, based on the date of visit, is shown in Figure 10c. Ignoring the colour, we observe prediction drifts around the year 2012 and around the year 2018, with a sharp drop in the risk of progression. This could be related to a more widespread use of high-efficacy disease-modifying therapies (see also Section D.2) and, more generally, to improvements in the clinical code of practice. The interaction given by the colour shows that the SHAP values of the date of visit correlate the most with the maximum cerebellar KFS score over the past 3 years. For old visits, high cerebellar KFS scores increase the risk of progression, while for more recent visits (i.e., from around 2016) the opposite is true. Note, however, that this can not be attributed to a big shift in the disease duration of the population over time, as that has remained mostly constant.[19]

4.3. Local explanations

In this section, we look at local (i.e., visit-specific) explanations of RSF predictions, generated by SHAP and Bellatrex. Explanations are generated for two baseline visits with differing prognostic outcomes. The trajectories of these patients, with the baseline visit (t_0) indicated, are shown in Figure 11. The first baseline visit has a favourable outcome (no progression in the next 4 years): a Canadian clinic following an approximately 50-year-old patient that was diagnosed with RRMS in the early 2000s, 11 years before the visit of interest (t_0). The second visit has an unfavourable outcome (progression within 2 years): an Australian clinic following an approximately 30-year-old patient that was diagnosed with RRMS in the early 2000s, 6 years before the visit of interest (t_0). To get an idea of how the predicted progression at the time of the visit of these patients compares with the general population, we compare their estimated survival curve to the Kaplan-Meier estimate over the training data in Figure 12.

[18]The Spearman rank correlation coefficient between max EDSS and disease duration is moderately positive at 0.38 (and statistically significant, $p = 0$ for $n = 321,775$).

[19]The Spearman rank correlation coefficient between the date of visit and disease duration is rather small at 0.092. There is a slight positive correlation: more recent dates of visit thus have slightly higher disease duration, due to existing patients continuously getting older. However, this is mostly counterbalanced by new patients (early in their disease duration) entering the cohort.

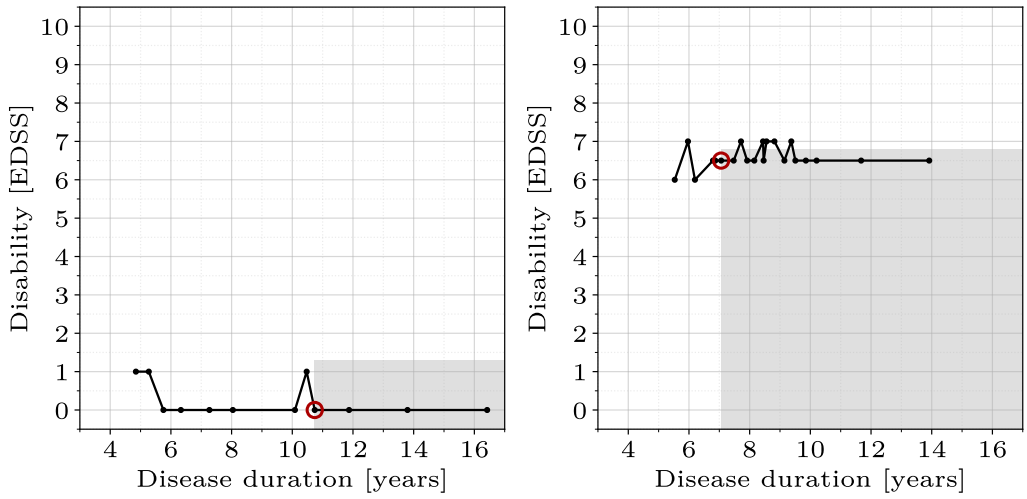


Figure 11: Low-risk (left) and high-risk (right) patient visit, each with the patient trajectory as context. The gray region indicates the zone of non-significant progression (with significance of progression defined in Section 3.2).

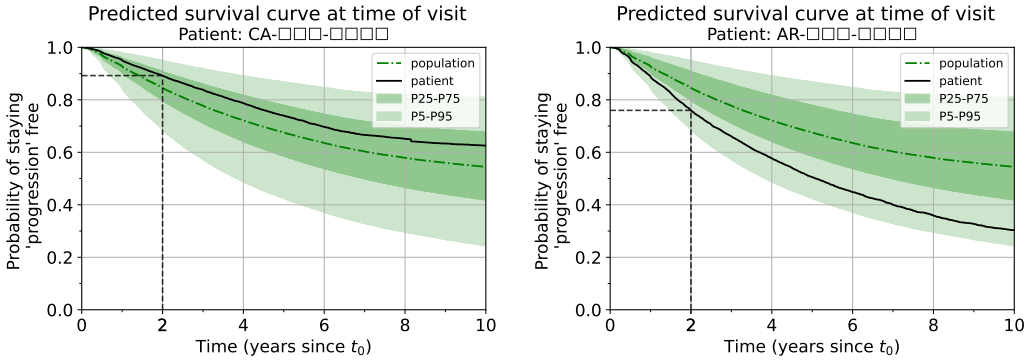


Figure 12: Predicted survival curve at the time of visit t_0 of a low-risk patient visit (left) and a high-risk patient visit (right) compared to the general population (bands indicating the percentile ranges). Note that the low-risk patient is predicted to worsen within 2 years with a probability of $\approx 10\%$ (= 1.0 - 0.9), whereas the high-risk patient with $\approx 25\%$ (= 1.0 - 0.75).

SHAP local explanations for the two patients are provided in Figure 13, where we see how, starting from a baseline prediction, the input features contribute to the model prediction. Feature values that drive the prediction to the right (red arrows) are linked to increased risk, whereas feature values associated with a negative contribution (blue arrows) are associated with lower risk.

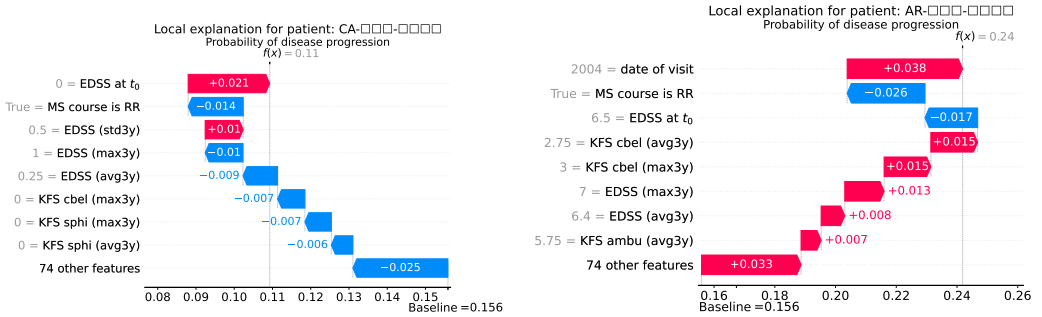


Figure 13: Local SHAP explanation for a low-risk patient visit (left) and a high-risk patient visit (right) for the time-to-event prediction by the RSF model.

The low-risk patient is predicted to worsen with a probability $\hat{p}_i$ of 10.9% over the next two years, compared to a population baseline of 15.6%. The explanation of the output for such low-risk patient shows similar insights compared to the global explanations identified in Figure 9b: a low average EDSS score and KFS 2 is associated with a better prognosis, and so is a more recent date of visit. Interestingly, a current EDSS score 0 coupled with a higher maximum EDSS score (equal to 1 in this case) is associated with a worse outcome, likely due to a regression to the mean effect. This effect suggests that the initial improvement to EDSS score 0 might be part of natural variability in the disease progression (or even measurement error), and that over time the patient's EDSS score is likely to revert back towards their average level of disability. See also the discussion about EDSS and maximum or average EDSS in Section 4.2.1 with the reference examples from Figure 3.

On the other hand, the high-risk patient is predicted to experience progression with a probability $\hat{p}_i$ of 24.2%. The strongest drivers of this prediction are a relatively high cerebellar KFS score over the past 3 years and a relatively old date of visit. Similarly to the low-risk patient, the contribution of the maximum value of the EDSS score has an opposite sign with respect to the

current EDSS score, this time in the other direction (regression to the mean, but increasing)

These explanations align well with the current knowledge of RRMS patients having a better EDSS trajectory on average than progressive MS disease courses [1], whereas the better prognosis for more recent visits can be associated with improved healthcare quality for MS [30] (see Section 4.2.1 for more details). To enhance and validate the findings given by SHAP explanations, we now make use of Bellatrex. This toolbox offers additional perspectives, ensuring a more comprehensive and robust analysis of our model's decision-making on a local scale. The resulting explanations are shown in Figure 14. The interpretation of these explanations is similar to SHAP: rules that are driving the prediction towards the right are associated with an increased risk, whereas rules driving the prediction to the left are associated with a decreased risk.

The explanations provided by the rule extraction process of Bellatrex confirm the previous findings. For the low-risk patient, the MS course is driving the predicted risk downwards, together with a relatively recent visit date. Furthermore, relatively low cerebellar and sphincteric KFS scores are also contributing towards a better prognosis. The high-risk patient has a similar explanation pattern, with the MS course driving the predicted risk downwards and a relatively high past EDSS score driving the predicted risk upwards.

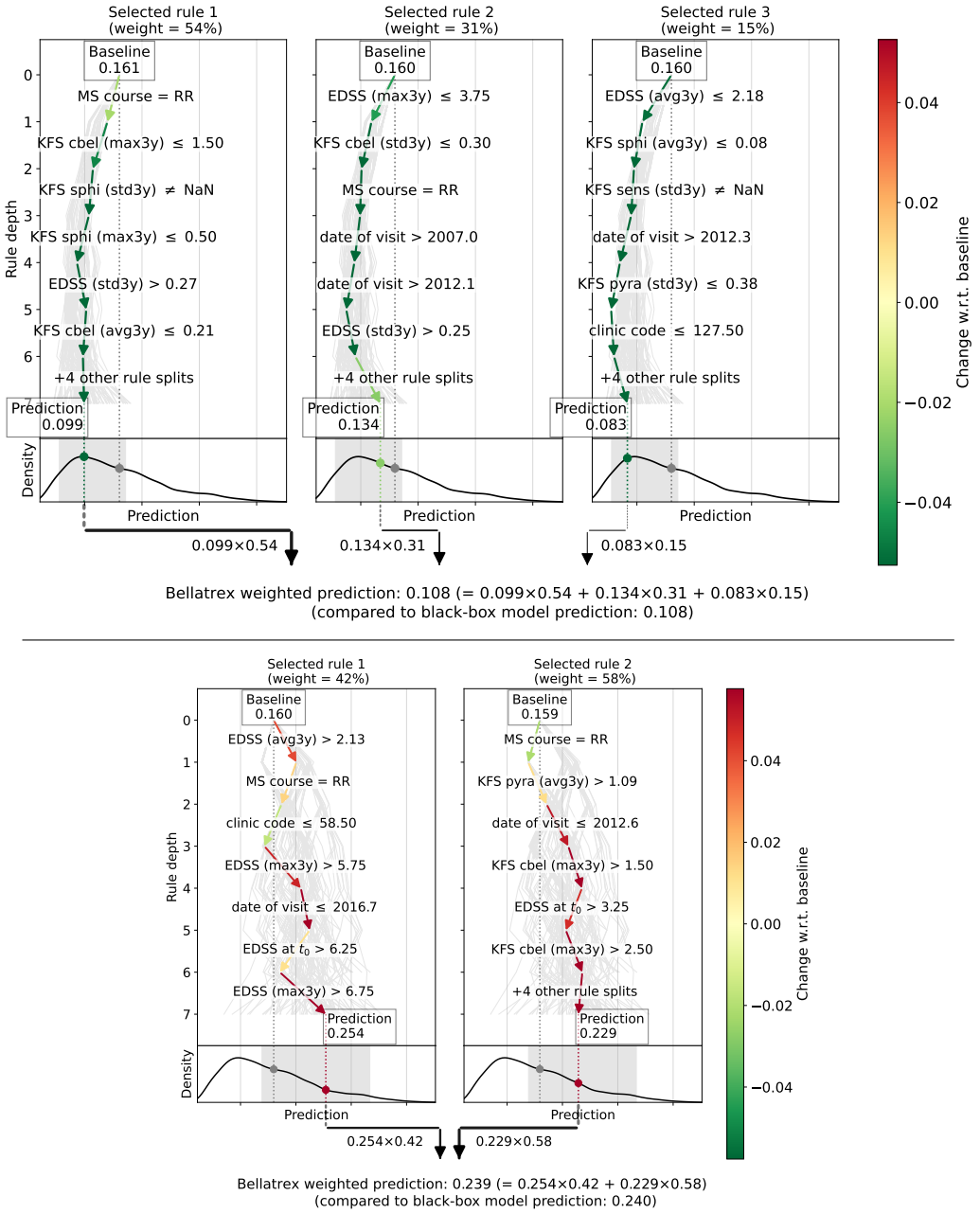


Figure 14: Local Bellatrex explanation for a low-risk patient visit (top) and a high-risk patient visit (bottom) for the time-to-event prediction by the RSF model. The grey area at the bottom of each plot represents the 90% confidence interval of the model's predictions. This interval is determined by identifying the range within which 90% of the predictions from the individual tree estimators fall.

5. Discussion

In this study, we explored the predictive performance and interpretability of time-to-progression machine learning models applied to the international MSBase registry, with a particular focus on the Random Survival Forest (RSF) model. The study of interpretability has two main target audiences: the machine learning (ML) engineer and the clinician itself. Whereas current computerised decision support systems are mostly informing the user about performance metrics, our work with explainability techniques showcases how the end user experience could in the future be further enhanced. For the ML engineer, graphs such as Figure 14 could be used to debug a predictive model (in collaboration with the clinical domain expert). For the clinician, graphs such as Figure 13 could ameliorate current patient dashboards and provide insights into patient prognosis, as well as increase trust in predictions generated by a black-box model. The main next step for this work is to improve predictive performance, and is discussed further in this section as a limitation of the current study.

A primary strength of this work is the capacity for making predictions across a wide time horizon (up to 15 years were considered here), offering a comprehensive longitudinal view of patient outcomes. This ability is crucial in a long-course disease like MS, where understanding progression over time is essential for effective treatment planning. Furthermore, the methodology employed addresses the issue of label bias encountered in the standard binary setting where 'progression in X years' is predicted. In particular, relatively recent visits can now more safely be included: as opposed to merely labelling these as 'no progression' as in the binary classification set-up (or dropping them in case of progression without confirmation available), they are labelled as 'censored' relatively early in time. This provides a more nuanced target variable and enhances the reliability of the predictions. Finally, the explanations provided by the interpretability tools, SHAP and Bellatrex, are another noteworthy strength of this study. Their outputs are consistent with current knowledge and therefore can offer valuable insights to clinical experts during their decision-making processes. In future work, a user study can be conducted in collaboration with clinicians to further validate and improve the interpretability.

Our findings also reveal several insights, which we discuss in the context of previous literature. Firstly, our methodology achieves a predictive performance that is consistent with previous work [15], despite the implementation

of different methodological approaches. This similarity suggests that we have reached a performance ceiling with the current predictive ML modelling approach on the MSBase dataset, likely due to the intrinsic noise in EDSS measurements [33]. Additionally, the fact that model performance remains mostly unchanged, regardless of whether train-test folds are grouped by clinic (the main setup) or by patient, is particularly striking. This is unexpected because different clinics often employ varied methods and standards of care, broadly influencing both the input data and the output labels of the dataset. Normally, such variations would lead to significant performance disparities, but this was not observed in our analysis. This highlights once more the performance ceiling reached with the current predictive ML modelling approach on the MSBase dataset.

Secondly, when local explanations are generated, SHAP and Bellatrex offer distinct types of explanations: point-wise SHAP explanations refer to specific feature values, whereas rule-based explanations for Bellatrex focus on feature thresholds. Crucially, these explanations show consistent trends. Such alignment not only validates the robustness of the findings but also provides a more complete view of the model's behaviour, enhancing interpretability through complementary perspectives.

Finally, given the underlying international multi-cohort nature of the MSBase dataset, we extend the applicability of the findings beyond individual clinics or regions, similarly to the work performed in [15]. This global perspective is reinforced by the validation of model performance using what effectively serves as an external cohort (test data comes from disjoint clinics with respect to the training data), further attesting to the robustness and generalisability of the study's outcomes.

Despite its considerable strengths, this study is not without limitations. One of the most significant is the limited performance of the predictive models, which currently do not meet the threshold for practical application in clinical settings. This performance is reflective of the heterogeneity of the disease course and of the limitations in using EDSS scores to define disability progression, despite their widespread use. Nevertheless, this gap between research findings and clinical utility highlights a crucial area for future improvement. Performance could potentially be improved by using more data modalities (e.g., magnetic resonance imaging (MRI) and genetic markers), offering a more comprehensive view of patient status.

Another limitation is related to the portability of the SHAP explanations. Normally, one needs access to the complete training dataset to generate

explanations with SHAP. However, data availability and privacy concerns can limit the feasibility of making such analyses in clinical environments. Fortunately, for models like RF and RSF an approximation of the SHAP values can be made based only on the model itself, not needing to make use of the data. To showcase how the SHAP explanations would be generated in practice, this approximation has been used throughout this work. However, this approximation comes at the cost of reduced accuracy of the obtained SHAP values. Thus, there is always a portability-performance trade-off to be made with respect to generating SHAP values.

6. Conclusions

In this research, we employed machine learning (ML)-driven and interpretable time-to-progression predictions in multiple sclerosis (MS). More specifically, we utilised Random Survival Forest (RSF) for predicting time to confirmed disease progression, and we enriched model predictions with post-hoc explanations generated by SHAP [27] and Bellatrex [29]. The modelling and the explanations were performed on an international, multi-centre dataset for research on MS, namely the MSBase [23] registry. The results were compared against existing state-of-the-art algorithms trained on the same data, that focus on binary predictions at a fixed time horizon (2 years) instead, and we demonstrated we can achieve comparable performance when we limit our model predictions to this same fixed time horizon. We focused our comparison on the Random Forest model used by [16], ensuring not only a comparison between models with similar architectures, but also the use of the same interpretability toolboxes. This strategy allowed us to narrow down the causes of the differences in model behaviour solely to the type of data and inclusion criteria.

Our contributions to the field are multifaceted and are marked by the following advancements. Firstly, we developed a machine learning-based time-to-event model that sets a new benchmark for long-horizon predictions, demonstrating state-of-the-art performance in survival analysis. Thanks to the rich patient follow-up present in MSBase, this model was able to provide a comprehensive perspective on the longitudinal patient trajectory. Secondly, in a broader context, we set up a pipeline designed to harness the full potential of time-to-event models. The novel pipeline approach allows the inclusion of more recent visits without the need to discard data. Lastly, we used and compared two Explainable AI toolboxes: SHAP [27] and Bellatrex [29],

which have been instrumental in shedding light on the inner workings of our models. Thanks to these techniques, we were able to detect a reduced effect of artefacts or spurious correlations created by combination of the proposed model and inclusion criteria. Additionally, we have found predictive patterns that are compatible with clinical insights, ranging from the importance of the KFS cerebellar and pyramidal score, to the importance of early use of highly-effective disease modifying therapies.

Within the context of our contributions, it is important to provide realistic expectations about the direct impact of our findings. The RSF model's capacity for time-to-event predictions over a long time horizon offers promising insights. Yet, it is crucial to acknowledge that the current performance level is insufficient for direct application in clinical settings. Additionally, the inclusion of the date of visit as a predictor only has a retrospective value, and should not be extrapolated beyond the current range of the training data (September 2020) to avoid bias. Our research serves more as a proof of concept, illustrating the potential of such models to enhance clinical decision-making processes once data quality is improved. This could include the integration of MRI data or other relevant clinical metrics and, at the same time, promote a more meticulous and standardised data collection procedure by increasingly data-literate clinicians and nurses.

Declarations and statements

Declaration of competing interest

- Tomas Kalincik served on scientific advisory boards for MS International Federation and World Health Organisation, BMS, Roche, Janssen, Sanofi Genzyme, Novartis, Merck and Biogen, steering committee for Brain Atrophy Initiative by Sanofi Genzyme, received conference travel support and/or speaker honoraria from WebMD Global, Eisai, Novartis, Biogen, Roche, Sanofi-Genzyme, Teva, BioCSL and Merck and received research or educational event support from Biogen, Novartis, Genzyme, Roche, Celgene and Merck.
- Eva Kubala Havrdova received honoraria/research support from Biogen, Merck Serono, Novars, Roche, and Teva; has been member of advisory boards for Actelion, Biogen, Celgene, Merck Serono, Novars, and Sanofi Genzyme;received honoraria/research support from Biogen, Merck Serono, Novars, Roche, and Teva; has been member of advisory boards for Actelion, Biogen, Celgene, Merck Serono, Novars, and Sanofi Genzyme; and has been

supported by the Czech Ministry of Education – project Cooperatio LF1, research area Neuroscience, and the project National Institute for Neurological Research (Programme EXCELES, ID project No LX22NPO5107) – funded by the European Union-Next Generation EU.

- Alessandra Lugaresi has received personal compensation for consulting, serving on a scientific advisory board, speaking or other activities from Alexion, Biogen, Bristol Myers Squibb, Horizon, Janssen, Merck Serono, Novartis, and Sanofi/Genzyme, and Her institutions have received research grants from Novartis and Sanofi/Genzyme.
- Bianca Weinstock-Guttman served as a consultant for Biogen, EMD Serono, Novartis, Genentech, Celgene/Bristol Meyers Squibb , Sanofi Genzyme, Bayer, Janssen, Labcorp , Horizon and SANA. Dr. Weinstock-Guttman also has received grant/research support from Novartis, Biogen, Horizon/Amgen. She serves in the editorial board for Children, CNS Drugs, MS International, Journal of Neurology Frontiers Epidemiology.
- Saloua Mrabet has received a MENACTRIMS clinical fellowship grant (2020).
- Patrice Lalive received honoraria for speaking and or travel expense from Biogen, Merck, Novartis, Roche; consulting fees from Biogen, GeNeuro, Merck, Novartis, Roche; research support from Biogen, Merck, Novartis. None were related to this work.
- Allan G Kermode received speaker honoraria and scientific advisory board fees from Bayer, BioCSL, Biogen, Genzyme, Innate Immunotherapeutics, Merck , Novartis, Sanofi, Sanofi-Aventis, and Teva.
- Stephen Reddel has received funds over the last 5 years including but not limited to travel support, honoraria, trial payments, research and clinical support to the institution from Alexion, Biogen, Merck, Novartis, Roche, Sandoz, Sanofi. Additional interests and potential conflicts of interest include: Australian Technical Advisory Group on Immunisation Varicella Zoster working party (unpaid).
- Francesco Patti received personal compensation for serving on advisory board by Almirall, Alexion, Biogen, Bristol, Janssen, Merck, Novartis and Roche. He further received research grant by Alexion, Almirall, Biogen, Bristol, Merck, Novartis and Roche and by FISM, Reload Association (Onlus), Italian Health Minister, and University of Catania.
- Valentina Tomassini has received consultation and speaker fees, travel grants and research support from: Biogen, Sanofi Genzyme, Merck, Novartis, Roche, Alexion, Viatris, Janssen, Bristol Myers Squibb, Almirall.

- Izanne Roos has served on scientific advisory boards, received conference travel support and/or speaker honoraria from Roche, Novartis, Merck and Biogen. Izanne Roos is supported by a MS Australia and the Trish Multiple Sclerosis Research Foundation.
- Raed Alroughani received honoraria as a speaker and for serving on scientific advisory boards from Bayer, Biogen, GSK, Merck, Novartis, Roche and Sanofi-Genzyme.
- Samia J. Khoury received compensation for scientific advisory board activity from Merck and Roche, and received compensation for serving on the IDMC for Biogen.
- Vincent van Pesch received travel grants from Merck Healthcare KGaA (Darmstadt, Germany), Biogen, Sanofi, Bristol Meyer Squibb, Almirall and Roche. His institution has received research grants and consultancy fees from Roche, Biogen, Sanofi, Merck Healthcare KGaA (Darmstadt, Germany), Bristol Meyer Squibb, Janssen, Almirall, Novartis Pharma, and Alexion.
- Maria Jose Sa received consulting fees, speaker honoraria, and/or travel expenses for scientific meetings from Alexion, Bayer Healthcare, Biogen, Bristol Myers Squibb, Celgene, Janssen, Merck-Serono, Novartis, Roche, Sanofi and Teva.
- Julie Prevost accepted travel compensation from Novartis, Biogen, Genzyme, Teva, and speaking honoraria from Biogen, Novartis, Genzyme and Teva.
- Daniele Spitaleri received honoraria as a consultant on scientific advisory boards by Bayer-Schering, Novartis and Sanofi-Aventis and compensation for travel from Novartis, Biogen, Sanofi Aventis, Teva and Merck.
- Pamela McCombe received speakers fees and travel grants from Novartis, Biogen, T'évalua, Sanofi.
- Claudio Solaro served on scientific advisory boards for Merck, Genzyme, Almirall, and Biogen; received honoraria and travel grants from Sanofi Aventis, Novartis, Biogen, Merck, Genzyme and Teva.
- Anneke van der Walt served on advisory boards and receives unrestricted research grants from Novartis, Biogen, Merck and Roche She has received speaker's honoraria and travel support from Novartis, Roche, and Merck. She receives grant support from the National Health and Medical Research Council of Australia and MS Research Australia.
- Helmut Butzkueven received institutional (Monash University) funding from Biogen, Roche, Merck, Alexion and Novartis; has carried out con-

tracted research for Novartis, Merck, Roche and Biogen; has taken part in speakers' bureaus for Biogen, Novartis, Roche and Merck; has received personal compensation from Oxford Health Policy Forum for the Brain Health Steering Committee.

- Guy Laureys received travel and/or consultancy compensation from Sanofi-Genzyme, Roche, Teva, Merck, Novartis, Celgene, Biogen.
- Jose Luis Sanchez-Menoyo accepted travel compensation from Novartis, Merck and Biogen, speaking honoraria from Biogen, Novartis, Sanofi, Merck, Almirall, Bayer and Teva and has participated in clinical trials by Biogen, Merck and Roche.
- Koen de Gans served on scientific advisory boards for Roche, Janssen, Sanofi-Genzyme, Novartis and Merck, received conference fee and travel support from Novartis, Biogen, Sanofi-Genzyme, Teva, Abbvie and Merck and received educational event support from Novartis.
- Abdullah Al-Asmi received personal compensation for serving as a Scientific Advisory or speaker/moderator for Novartis, Biogen, Roche, Sanofi-Genzyme, and Merck.
- Norma Deri received funding from Bayer, Merck , Biogen, Genzyme and Novartis.
- Tunde Csepany received speaker honoraria/ conference travel support from Biogen, Merck, Novartis, Roche, Sanofi-Aventis and Teva.
- William M Carroll received travel assistance and honoraria for participation in industry sponsored meetings from and provided advice to Bayer Schering Pharma, Biogen-Idec, Novartis, Roche, Genzyme, Sanofi-Aventis, CSL, Teva, Merck and Cellgene.
- Csilla Rozsa received speaker honoraria from Bayer Schering, Novartis and Biogen, congress and travel expense compensations from Biogen, Teva, Merck and Bayer Schering.
- Bhim Singhal received consultancy honoraria and compensation for travel from Biogen and Merck.
- Todd A. Hardy received speaker honoraria/ conference travel support or served on advisory boards for Bayer Schering, Biogen, Merck , Novartis, Roche, Sanofi-Genzyme, Bristol Myers Squibb and Teva.
- Sudarshini Ramanathan has received research funding from the National Health and Medical Research Council (NHMRC, Australia), the Petre Foundation, the Brain Foundation, the Royal Australasian College of Physicians, and the University of Sydney. She is supported by an NHMRC Investigator Grant (GNT2008339). She serves as a consultant on the Inter-

national Steering Committee for a clinical trial led by UCB (NCT05063162). She is on the advisory board for educational activities led by Limbic Neurology. She has been an invited speaker for educational/research sessions coordinated by Biogen, Alexion, Novartis, Excemed and Limbic Neurology. She is on the medical advisory board (non-remunerated positions) of The MOG Project and the Sumaira Foundation.

Author CRediT statement

- **Conceptualisation**: Robbe D'hondt, Klest Dedja, Sofie Aerts, Liesbet Peeters, Celine Vens,
- **Data curation**: Robbe D'hondt, Klest Dedja, the MSBase Study Group*,
- **Formal analysis**: Robbe D'hondt, Klest Dedja,
- **Funding acquisition**: Robbe D'hondt (Research Foundation Flanders grant), Celine Vens (Research Foundation Flanders grant, Flemish government AI Research Program),
- **Investigation**: the MSBase Study Group*,
- **Methodology**: Robbe D'hondt, Klest Dedja, Celine Vens,
- **Project administration**: Robbe D'hondt, Klest Dedja, Liesbet Peeters, Celine Vens,
- **Resources**: Celine Vens (computing resources),
- **Software**: Robbe D'hondt, Klest Dedja,
- **Supervision**: Liesbet Peeters, Celine Vens,
- **Validation**: Sofie Aerts, Bart van Wijmeersch, Liesbet Peeters,
- **Visualisation**: Robbe D'hondt, Klest Dedja
- **Writing – original draft**: Robbe D'hondt, Klest Dedja
- **Writing – review and editing**: all authors,

*In the above statement, the following authors are part of, and referred to as, 'the MSBase Study Group': Bart Van Wijmeersch, Tomas Kalincik, Stephen Reddel, Eva Kubala Havrdova, Alessandra Lugaresi, Bianca Weinstock-Guttman, Saloua Mrabet, Patrice Lalive, Allan G Kermode, Serkan Ozakbas, Francesco Patti, Alexandre Prat, Valentina Tomassini, Izanne Roos, Raed Alroughani, Oliver Gerlach, Samia J. Khoury, Vincent van Pesch, Maria Jose Sa, Julie Prevost, Daniele Spitaleri, Pamela McCombe, Claudio Solaro, Anneke van der Walt, Helmut Butzkueven, Guy Laureys, Jose Luis Sanchez-Menoyo, Koen de Gans, Abdullah Al-Asmi, Norma Deri, Tunde Csepany, Talal Al-Harbi, William M Carroll, Csilla Rozsa, Bhim Singhal, Todd A. Hardy, Sudarshini Ramanathan.

Acknowledgments

Funding: This work was supported by Research Foundation Flanders [grant number 1S38023N] and the Flemish government AI Research Program (FAIR).

Contributors: We acknowledge data contributions from the following MSBase principal investigators, ordered by the number of contributed patients (from high to low): Dana Horakova, Guillermo Izquierdo, Sara Eichau, Marc Girard, Pierre Duquette, Pierre Grammond, Francois Grand'Maison, Maria Pia Amato, Katherine Buzzard, Cavit Boz, Murat Terzi, Vahid Shaygannejad, Jeannette Lechner-Scott, Jens Kuhle, Bassem Yamout, Yolanda Blanco, Elisabetta Cartechini, Recai Turkoglu, Nevin John, Radek Ampapa, Davide Maimone, Cristina Ramo-Tello, Celia Oreja-Guevara, Maria Di Gregorio, Mark Slee, Aysun Soysal, Riadh Gouider, Richard Macdonell, Maria Edite Rio, Liesbeth Van Hijfte, Jiwon Oh, Tamara Castillo-Triviño, Michael Barnett, Ricardo Fernandez Bolaños, Marie D'hooghe, Justin Garber, Ayse Altintas, Cees Zwanikken, Eduardo Aguera-Morales, Magd Zakaria, Sarah Besora, Suzanne Hodgkinson, the late Yara Fragoso, Rana Karabudak, Edgardo Cristiano, Jose Antonio Cabrera-Gomez, Maria Laura Saladino, Leontien Den braber-Moerland, Bruce Taylor, Orla Gray, Shlomo Flechter, Fraser Moore, Claudio Gobbi, Chris McGuigan , Jennifer Massey, Jamie Campbell, Marzena Fabis-Pedrini, Nevin Shalaby, Mihaela Simu, Angel Perez sempere, Cameron Shaw, Jan Schepel, Steve Vucic, Jabir Alkhaboori, Magdolna Simo, Danny Decoo, Jose Andres Dominguez, Neil Shuey, Stella Hughes, Ilya Kister.

Finally, we acknowledge for their aid in data acquisition: Dr Mark Marriott, Dr Trevor Kilpatrick, Dr John King, Dr Katherine Buzzard, Dr Ai-Lan Nguyen, Dr Chris Dwyer, Dr Mastura Monif, Dr Izanne Roos, Ms Lisa Taylor, Ms Josephine Baker, Prof Robert Zivadinov, Prof Ralph Benedict, Dr Marzena Fabis-Pedrini, Dr Clara Chisari, Dr Emanuele D'Amico, Dr Lo Fermo Salvatore, Dr Catherine Larochelle, Dr Raymond Hupperts, Dr Freek Verheul, Dr Krisztian Kasa.

Data availability statement

Data is available upon reasonable request. The data is the property of the individual centres. Data from the participating cohorts can be requested from the principal investigators, at their discretion and conditional on obtaining approvals from the appropriate institutional review boards. The MSBase Registry is a data processor and warehouses data from individual principal investigators who agree to share their datasets on a project-by-project basis.

Data access to external parties can be granted on reasonable request at the sole discretion of the principal investigators, who will need to be approached individually for permission.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

A. Data dictionary

Below, we provide a list of all 82 input features used for modelling, with their names and an explanation of the meaning.

MS course : Three binary variables indicating whether the MS disease course is **RR** (relapsing-remitting), **PP** (primary progressive) or **SP** (secondary progressive).

Education : Three binary variables indicating whether the education status of the patient is **higher**, **lower**, or **unknown**.

Sex : A binary variable indicating the biological sex of the patient assigned at birth (female: 1 or male: 0).

country code, clinic code : Index of the country and clinic in an alphabetically sorted list of all countries and clinics.

date of visit, onset date : The date of t_0 and of MS onset.

age, age at onset : The age of the patient at t_0 and at MS onset.

disease duration : Time between MS onset and t_0.

First symptom : Four binary variables indicating whether the first MS symptoms for this patient were related to the **optic pathways**, the **brainstem**, the **spinal cord** and/or the **supratentorial** region in the brain. Note that first symptoms can span multiple regions.

DMT, DMT IND : Two binary variables indicating whether the patient is currently on a DMT (disease-modifying therapy) and/or an induction DMT respectively.

Number visits 3y : Number of visits for this patient in the past 3 years.

EDSS (...3y) : 6 disease history summary statistic variables: the **first**, **last**, **min** (minimum), **max** (maximum), **avg** (average) and **std** (standard deviation) EDSS score recorded over the past 3 years. For convenience, 'EDSS (last3y)' has been renamed to '**EDSS at t_0**'.

KFS ... (...3y) : 48 disease history summary statistic variables: same as above, but now for each of the eight KFS (Kurtzke Functional System) scores: **pyra** (pyramidal), **cbel** (cerebellar), **brai** (brainstem), **sens** (sensory), **sphi** (sphincteric), **visu** (visual), **cbra** (cerebral), and **ambu** (ambulatory).

relapse rate : Average number of yearly relapses at t_0 since first visit.

t since relapse : Number from 0 to 1 indicating propensity to relapses in the disease history of the patient. 1 indicates very recent relapse, 0 indicates no recorded relapses.

t since fampridine : Number from 0 to 1 indicating administration of fampridine in patient history, encoded in the same way as *t since relapse*.

t until treat : Number from 0 to 1 indicating treatment in patient history, encoded in the same way as *t since relapse*.

t until treat high : Number from 0 to 1 indicating high-efficacy treatment in patient history, encoded in the same way as *t since relapse*.

High DMT in past : A binary variable indicating whether this patient has at some point received a high-intensity DMT in its recorded trajectory before t_0.

ratio on treat : Fraction of the patient's history in which they were receiving a DMT.

B. Technical specifications

This appendix is dedicated to providing further technical details, and is aimed at the readers that are more familiar with machine learning. Our intention is to enhance the clarity of our work to the interested readers by delving into further details concerning the machine learning engineering point of view.

B.1. Binary label versus binarised label

As discussed in Appendix 3.4, the binary label employed in the in the 'progression in 2 years' framework is not equivalent to the binarised time-to-event label at $t = 2$. This non-equivalence also led to the AUROC and AUPRC being incomparable to their time-dependent counterparts (even for the same model), as observed in Appendix 4.1. Here, we elucidate the distinctions between the two labels.

The key point is to note that the visits censored before $t = 2$ are labelled as 'missing' by the binarisation process (Figure 15, right) but have a binary label 'false' (Figure 15, left) in the original binary classification set-up. This is because, in the binarisation process, we cannot place any generally valid assumptions on visits censored before t. We don't know if the event will occur at t or not: the actual label can still be either true or false. On the other hand, for the binary label we do know that all visits censored before 2 years are visits without progression, as the visits censored before 2 years *with* a progression event were already excluded from being valid baseline visits[20].

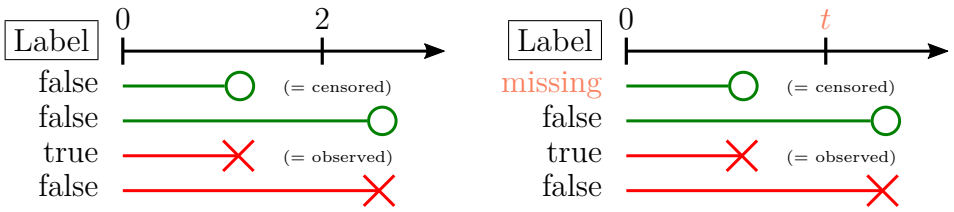


Figure 15: Left: relationship between the *binary* label from [15] (i.e., progression in 2 years) and possible survival analysis scenarios. Right: *binarisation* process of the survival analysis label to define time-dependent evaluation metrics.

B.2. Hyperparameter tuning

In Table 4, the hyperparameter grid used for machine learning model tuning is shown. To tune, we performed a randomized search for each model, with 10 hyperparameter combinations sampled at random from this grid. For more details about the machine learning training setup, please refer back to Appendix 3.3.

B.3. Alternative stratification scenario

Here we report the performance obtained in the alternative scenario where train-validation-test split stratification is performed based on patients rather than on clinics (reported in Appendix 4.1). The rationale for exploring this stratification approach arises from the expectation of obtaining increased performance outcomes, as the setting allows the models to learn about clinics-specific patterns that can be present in both training and testing data.

[20]The progression event can no longer be confirmed by another visit at least 2 years after $t = 0$. See also the definition of inclusion criteria for valid baseline visits in Appendix 3.1.

Table 4: Grid of possible hyperparameter combinations used for model tuning. p = number of input features (82 in our case).

Hyperparameter	**Random Forest**	**Random Survival Forest**
number of base learners	100	100
max depth	not limited	10
features considered per split	$\sqrt{p},\ p/2,\ p$	$\sqrt{p},\ p/2,\ p$
min weight fraction leaf	$0, 0.003, 0.01, 0.03, 0.1$	$0, 0.003, 0.01, 0.03, 0.1$

However, as shown in Table 5, there is minimal difference in performance between this stratification strategy and the group-by-clinic stratification in Table 2.

As a consequence, the focus of our paper remains on the clinic-based stratification, as it better serves as an external validation setting, and has the added advantage of being directly comparable to the settings in [15, 16].

Table 5: Performance results of the models trained on the binarised label, with folds that are grouped by patient. The average and standard deviation over the 5-folds are reported.

	AUROC	AUPRC
Random Forest	$0.702 \pm .010$	$0.240 \pm .018$
RSF restricted	$0.693 \pm .014$	$0.232 \pm .021$
RSF unrestricted	$0.696 \pm .011$	$0.236 \pm .020$

C. Comparison with other survival models

In Figure 16, we compare the time-dependent performance metrics for our RSF model to two alternative survival models (all on the augmented dataset). Specifically, we make the comparison to a regularized Cox proportional hazards model [20] and a survival Support Vector Machine (SVM) with a linear kernel [34]. The three methods behave pretty similar for short- and medium-term predictions, with RSF outperforming Cox and SVM in terms of AUROC when predicting long-term progression. We did not try other kernels for the SVM for computational considerations, as the kernel matrix needs over 400GB to be allocated.

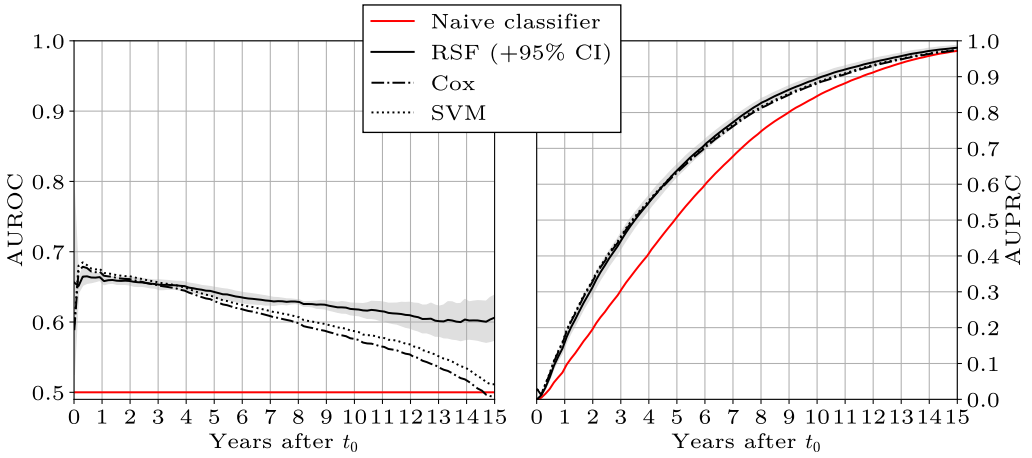


Figure 16: Time-dependent performance of RSF, regularized Cox, and SVM with linear kernel.

D. Additional data analysis

D.1. Performance on relapsing-remitting subgroup

In this appendix, we focus on the subgroup of relapsing-remitting patients. Since this subgroup has many different treatment options available, they would benefit the most from decision support systems for treatment switching. In Table 6, we report the performance results on the task of predicting progression in 2 years. Note that the AUROC is 2% lower than the general case. This is expected, as MSCOURSE was an important predictor in our dataset, so making a good ranking on a subset of the data that is homogeneous in this feature is more difficult than on the complete dataset. Ideally, the model should be retrained to get a more reliable performance estimate on this subset of patients, but that is outside of the scope of the current analysis.

Table 6: Performance results for predicting progression in 2 years for relapsing-remitting patients only. The average and standard deviation over the 5-fold cross-validation are reported. The naive classifier predicts the majority class for all baseline visits.

	AUROC	AUPRC
Naive classifier	0.500 ± 0.000	0.104 ± 0.007
RF	0.686 ± 0.005	0.186 ± 0.008
Restricted RSF	0.675 ± 0.012	0.176 ± 0.006
RSF	0.660 ± 0.009	0.172 ± 0.014

D.2. Use of high-efficacy treatments over the years

In Figure 17, we show how treatment administration has changed over time. In particular, we observe an increase in high-efficacy treatment around years 2008-2010 and 2018-2020. This justifies the remark in Appendix 4.2.2 regarding a potential explanation for the prediction drifts around those years.

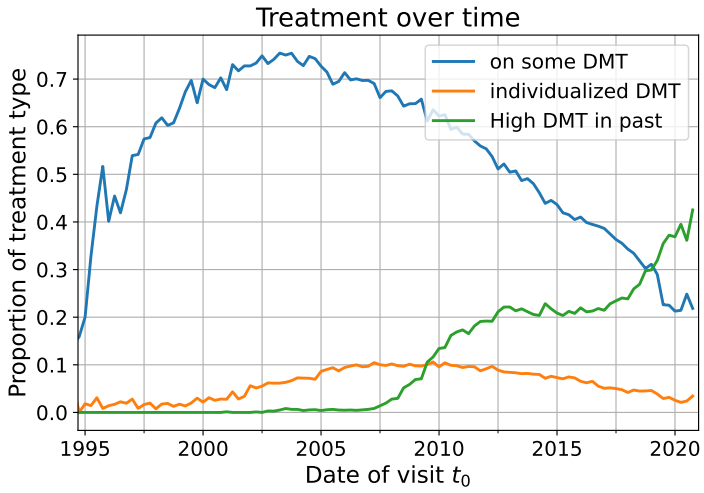


Figure 17: Evolution of DMT administration over time.

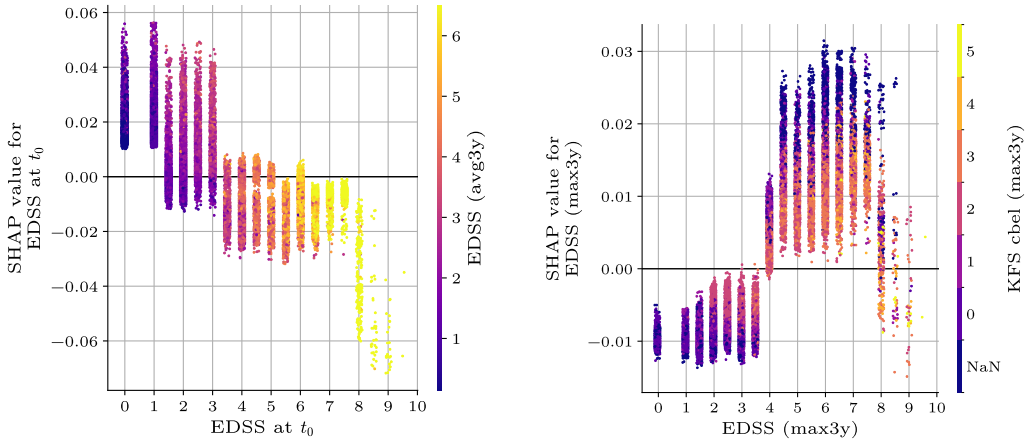
D.3. Interaction plots

In Figure 10, we have shown the SHAP interaction plots of three of the top five most predictive features for the RSF model. In Figure 18, we share the remaining two interaction plots, namely the one showing the strongest interaction with EDSS score (measured at time of baseline visit t_0) and the strongest interaction with maximum EDSS score (over the past 3 years before t_0).

D.4. Local interpretability

Here we clarify the criteria used to select patients from the test set to use as example explanation, and we provide the output explanations from SHAP and Bellatrex for four more selected patients.

We select 10 patients from the (first) test fold: 5 patients that are censored relatively late in time (60th, 65th, 70th, 75th and 80th percentile among all censored observations) representing good prognoses, and 5 patient with



(a) EDSS at t_0: SHAP values interact most with EDSS (max3y).

(b) EDSS (max3y): SHAP values interact most with KFS cbel (max3y).

Figure 18: Additional SHAP interaction plots.

confirmed progression within a relatively short period of time (30th, 35th, 40th, 45th and 50th percentile among all observed events) representing poor prognosis. Of these 10 patients, we discard the explanations where the model is either wrong (2 cases) or not confident enough (2 cases), leaving us with 6 cases. The two cases deemed most instructive are shown in the main text in Appendix 4.3, whereas here we share the explanations obtained by the remaining 4 patient visits.

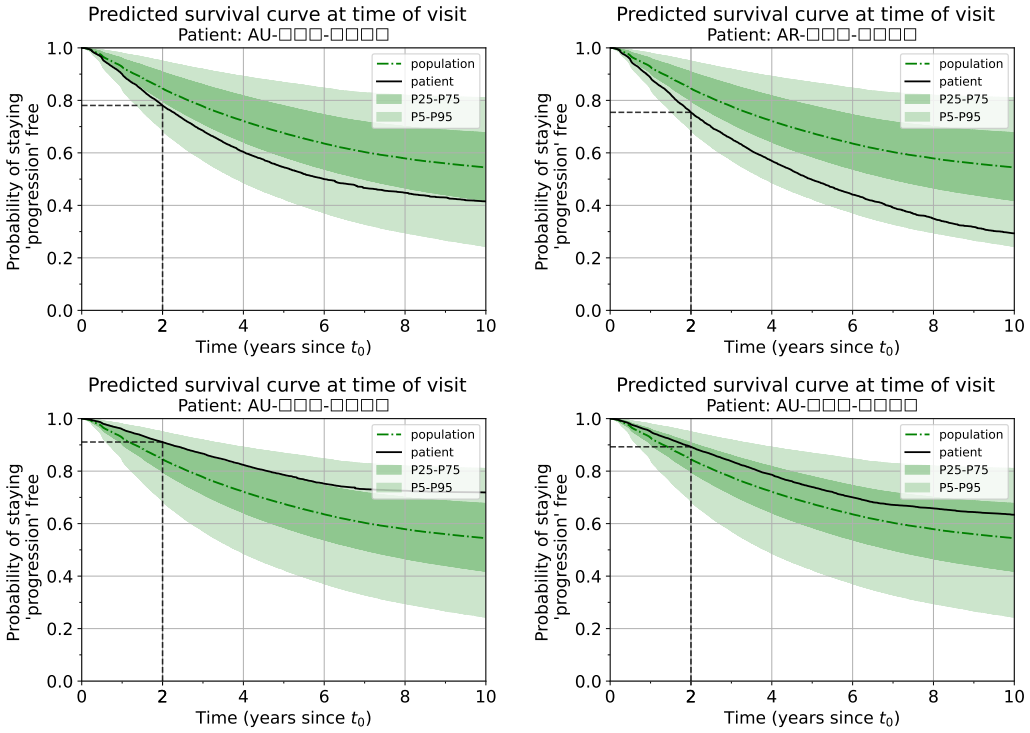


Figure 19: Predicted survival curve at time of visit t_0 of four patients, compared with the predictions for the visits in the training data.

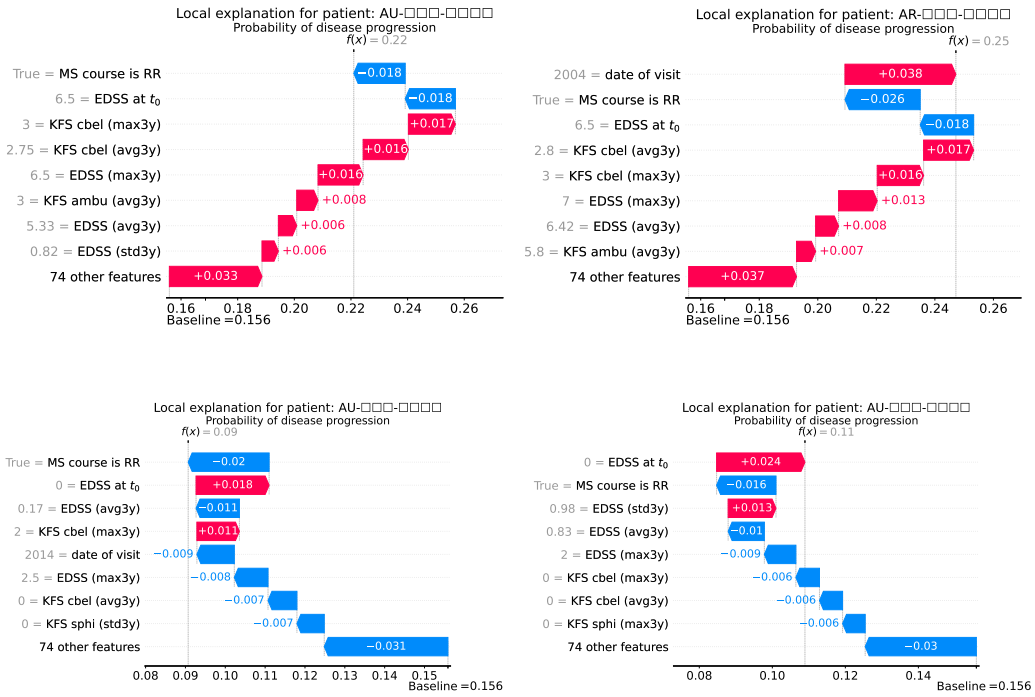
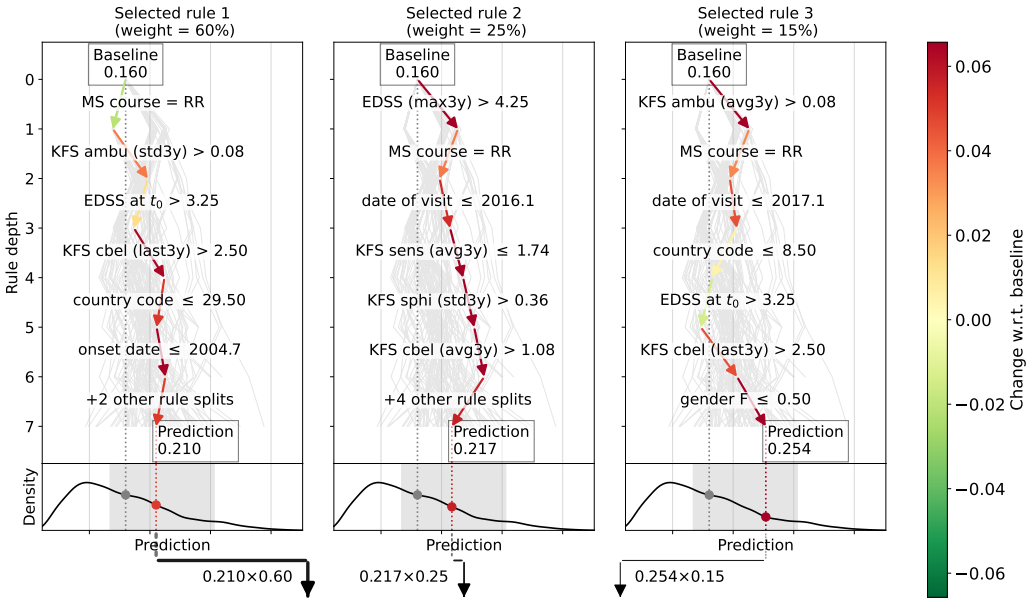
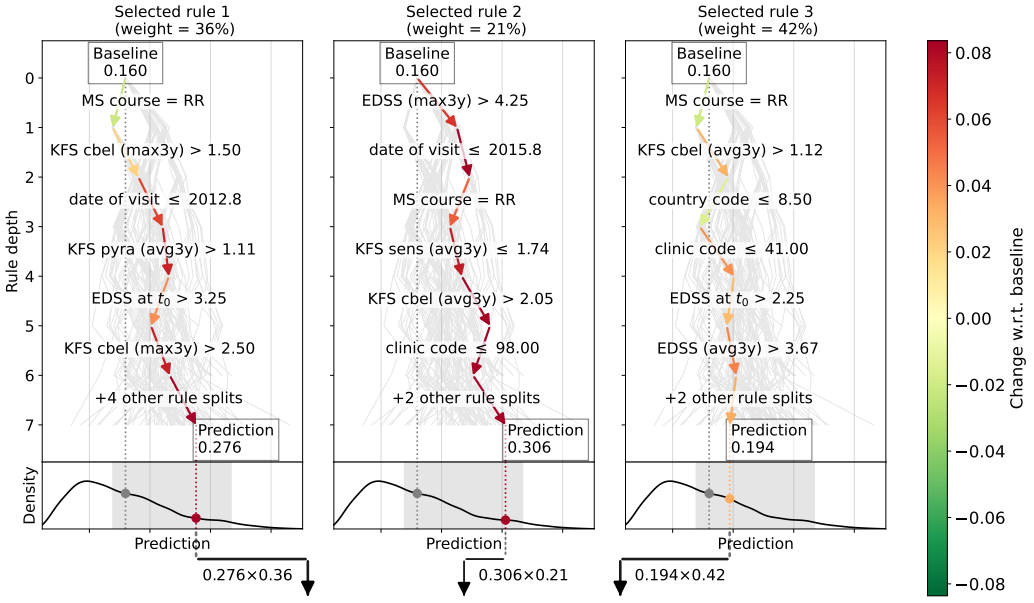


Figure 20: Local SHAP explanation for the RSF predictions made for two high-risk and two low-risk patient visits.



Bellatrex weighted prediction: 0.218 (= 0.210×0.60 + 0.217×0.25 + 0.254×0.15)
(compared to black-box model prediction: 0.219)



Bellatrex weighted prediction: 0.245 (= 0.276×0.36 + 0.306×0.21 + 0.194×0.42)
(compared to black-box model prediction: 0.246)

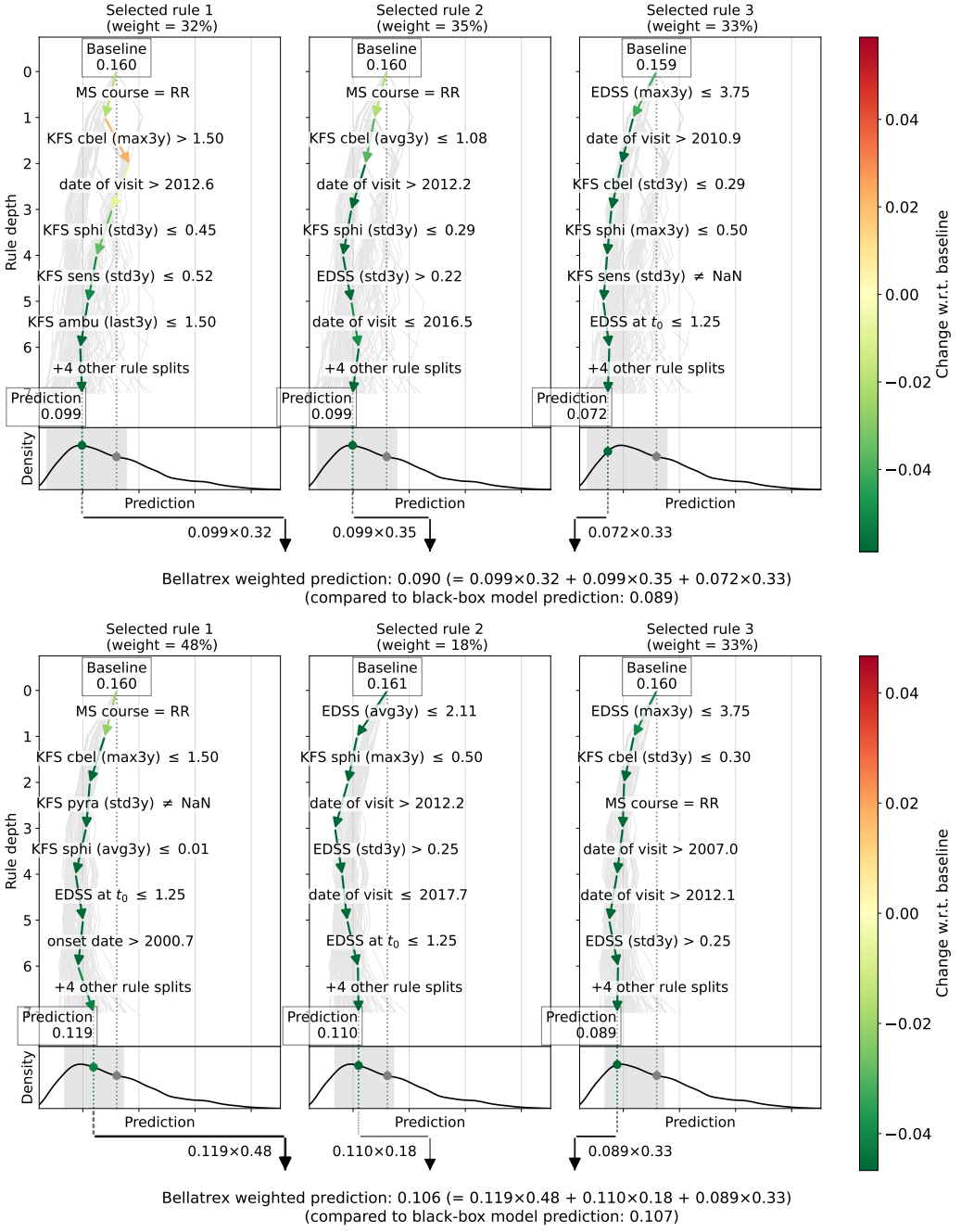


Figure 21: Local Bellatrex explanation for the RSF predictions made for two high-risk and two low-risk patient visits.

E. Implementation details

For reproducibility purposes, we share the software versions used for running the pipeline here: python 3.10, scikit-learn 1.3, scikit-survival 0.22, shap 0.42, pandas 1.5, pickle 4.0. It is worth noting that the dataset size and the resulting model size in the scikit-survival implementation make it challenging to run the pipeline on current laptops (see also https://github.com/sebp/scikit-survival/issues/343). The main authors have resorted, for the sake of time efficiency, into using a computing server with 16 cores (Intel Xeon Gold 6326) and 250 GB of RAM.

References

[1] M. Filippi, A. Bar-Or, F. Piehl, P. Preziosa, A. Solari, S. Vukusic, M. A. Rocca, Multiple sclerosis, Nat Rev Dis Primers 4 (1) (2018) 43. doi:10.1038/s41572-018-0041-4.

[2] J. F. Kurtzke, Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS), Neurology 33 (11) (1983) 1444–1452. doi:10.1212/wnl.33.11.1444.

[3] B. P. Çinar, Y. G. Yorgun, What We Learned from The History of Multiple Sclerosis Measurement: Expanded Disability Status Scale, Noro Psikiyatr Ars 55 (Suppl 1) (2018) S69–S75. doi:10.29399/npa.23343.

[4] M. Hartmann, N. Fenton, R. Dobson, Current review and next steps for artificial intelligence in multiple sclerosis risk research, Comput Biol Med 132 (2021) 104337. doi:10.1016/j.compbiomed.2021.104337.

[5] F. Moazami, A. Lefevre-Utile, C. Papaloukas, V. Soumelis, Machine learning approaches in study of multiple sclerosis disease through magnetic resonance images, Front Immunol 12 (2021) 3205. doi:10.3389/fimmu.2021.700582.

[6] R. Seccia, S. Romano, M. Salvetti, A. Crisanti, L. Palagi, F. Grassi, Machine learning use for prognostic purposes in multiple sclerosis, Life 11 (2) (2021) 122. doi:10.3390/life11020122.

[7] A. Balasundaram, M. K. Ghanta, Perspective Chapter: Artificial Intelligence in Multiple Sclerosis, in: Multiple Sclerosis - Genetics, Disease

Mechanisms and Clinical Developments, IntechOpen, 2023, Ch. 5, pp. 1–15. doi:10.5772/intechopen.113299.

[8] M. Trojano, M. Tintore, X. Montalban, J. Hillert, T. Kalincik, P. Iaffaldano, T. Spelman, M. P. Sormani, H. Butzkueven, Treatment decisions in multiple sclerosis — insights from real-world observational studies, Nat. Rev. Neurol. 13 (2) (2017) 105–118. doi:10.1038/nrneurol.2016.188.

[9] T. Kalincik, A. Manouchehrinia, L. Sobisek, V. Jokubaitis, T. Spelman, D. Horakova, E. Havrdova, M. Trojano, G. Izquierdo, A. Lugaresi, M. Girard, A. Prat, P. Duquette, P. Grammond, P. Sola, R. Hupperts, F. Grand'Maison, E. Pucci, C. Boz, R. Alroughani, V. Van Pesch, J. Lechner-Scott, M. Terzi, R. Bergamaschi, G. Iuliano, F. Granella, D. Spitaleri, V. Shaygannejad, C. Oreja-Guevara, M. Slee, R. Ampapa, F. Verheul, P. McCombe, J. Olascoaga, M. P. Amato, S. Vucic, S. Hodgkinson, C. Ramo-Tello, S. Flechter, E. Cristiano, C. Rozsa, F. Moore, J. Luis Sanchez-Menoyo, M. Laura Saladino, M. Barnett, J. Hillert, H. Butzkueven, Towards personalized therapy for multiple sclerosis: Prediction of individual treatment response, Brain 140 (9) (2017) 2426–2443. doi:10.1093/brain/awx185.

[10] A. Eshaghi, F. Prados, W. J. Brownlee, D. R. Altmann, C. Tur, M. J. Cardoso, F. De Angelis, S. H. van de Pavert, N. Cawley, N. De Stefano, M. L. Stromillo, M. Battaglini, S. Ruggieri, C. Gasperini, M. Filippi, M. A. Rocca, A. Rovira, J. Sastre-Garriga, H. Vrenken, C. E. Leurs, J. Killestein, L. Pirpamer, C. Enzinger, S. Ourselin, C. A. G. Wheeler-Kingshott, D. Chard, A. J. Thompson, D. C. Alexander, F. Barkhof, O. Ciccarelli, on behalf of the MAGNIMS study group, Deep gray matter volume loss drives disability worsening in multiple sclerosis, Ann Neurol 83 (2) (2018) 210–222. doi:10.1002/ana.25145.

[11] J.-P. R. Falet, J. Durso-Finley, B. Nichyporuk, J. Schroeter, F. Bovis, M.-P. Sormani, D. Precup, T. Arbel, D. L. Arnold, Estimating individual treatment effect on disability progression in multiple sclerosis using deep learning, Nat Commun 13 (1) (2022) 5645. doi:10.1038/s41467-022-33269-x.

[12] A. Eshaghi, P. A. Wijeratne, N. P. Oxtoby, D. L. Arnold, L. Collins, S. Narayanan, C. R. G. Guttmann, A. J. Thompson, D. C. Alexander,

F. Barkhof, D. Chard, O. Ciccarelli, Predicting personalised risk of disability worsening in multiple sclerosis with machine learning, medRxiv (Feb. 2022). doi:10.1101/2022.02.03.22270364.

[13] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, H. Aidos, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. M. Di Nunzio, P. Fariselli, J. M. García Dominguez, M. Gromicho, E. Longato, S. C. Madeira, U. Manera, G. Silvello, E. Tavazzi, E. Tavazzi, M. Vettoretti, B. Di Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2023, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 343–369. doi:10.1007/978-3-031-42448-9_24.

[14] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 623–631. doi:10.1145/2487575.2487579.

[15] E. De Brouwer, T. Becker, L. Werthen-Brabants, P. Dewulf, D. Iliadis, C. Dekeyser, G. Laureys, B. Van Wijmeersch, V. Popescu, T. Dhaene, D. Deschrijver, W. Waegeman, B. De Baets, M. Stock, D. Horakova, F. Patti, G. Izquierdo, S. Eichau, M. Girard, A. Prat, A. Lugaresi, P. Grammond, T. Kalincik, R. Alroughani, F. Grand'Maison, O. Skibina, M. Terzi, J. Lechner-Scott, O. Gerlach, S. J. Khoury, E. Cartechini, V. Van Pesch, M. J. Sà, B. Weinstock-Guttman, Y. Blanco, R. Ampapa, D. Spitaleri, C. Solaro, D. Maimone, A. Soysal, G. Iuliano, R. Gouider, T. Castillo-Triviño, J. L. Sánchez-Menoyo, G. Laureys, A. van der Walt, J. Oh, E. Aguera-Morales, A. Altintas, A. Al-Asmi, K. de Gans, Y. Fragoso, T. Csepany, S. Hodgkinson, N. Deri, T. Al-Harbi, B. Taylor, O. Gray, P. Lalive, C. Rozsa, C. McGuigan, A. Kermode, A. P. Sempere, S. Mihaela, M. Simo, T. Hardy, D. Decoo, S. Hughes, N. Grigoriadis, A. Sas, N. Vella, Y. Moreau, L. Peeters, Machine-learning-based prediction of disability progression in multiple sclerosis: An observational, international, multi-center study, PLOS Digital Health 3 (7) (2024) 1–25. doi:10.1371/journal.pdig.0000533.

[16] E. De Brouwer, T. Becker, Y. Moreau, E. K. Havrdova, M. Trojano,

S. Eichau, S. Ozakbas, M. Onofrj, P. Grammond, J. Kuhle, L. Kappos, P. Sola, E. Cartechini, J. Lechner-Scott, R. Alroughani, O. Gerlach, T. Kalincik, F. Granella, F. Grand'Maison, R. Bergamaschi, M. José Sá, B. Van Wijmeersch, A. Soysal, J. L. Sanchez-Menoyo, C. Solaro, C. Boz, G. Iuliano, K. Buzzard, E. Aguera-Morales, M. Terzi, T. C. Trivio, D. Spitaleri, V. Van Pesch, V. Shaygannejad, F. Moore, C. Oreja-Guevara, D. Maimone, R. Gouider, T. Csepany, C. Ramo-Tello, L. Peeters, Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression, Computer Methods and Programs in Biomedicine 208 (2021) 106180. doi:10.1016/j.cmpb.2021.106180.

[17] Y. Zhao, T. Wang, R. Bove, B. Cree, R. Henry, H. Lokhande, M. Polgar-Turcsanyi, M. Anderson, R. Bakshi, H. L. Weiner, T. Chitnis, SUMMIT Investigators, Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study, NPJ Digit Med 3 (2020) 135. doi:10.1038/s41746-020-00338-8.

[18] I. Dekker, A. J. C. Eijlers, V. Popescu, L. J. Balk, H. Vrenken, M. P. Wattjes, B. M. J. Uitdehaag, J. Killestein, J. J. G. Geurts, F. Barkhof, M. M. Schoonheim, Predicting clinical progression in multiple sclerosis after 6 and 12 years, Eur J Neurol 26 (6) (2019) 893–902. doi:10.1111/ene.13904.

[19] S. A. Gauthier, M. Mandel, C. R. G. Guttmann, B. I. Glanz, S. J. Khoury, R. A. Betensky, H. L. Weiner, Predicting short-term disability in multiple sclerosis, Neurology 68 (24) (2007) 2059–2065. doi:10.1212/01.wnl.0000264890.97479.b1.

[20] D. R. Cox, Regression models and life-tables, J Roy Stat Soc B Met 34 (1972) 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.

[21] J. Lorscheider, K. Buzzard, V. Jokubaitis, T. Spelman, E. Havrdova, D. Horakova, M. Trojano, G. Izquierdo, M. Girard, P. Duquette, A. Prat, A. Lugaresi, F. Grand'Maison, P. Grammond, R. Hupperts, R. Alroughani, P. Sola, C. Boz, E. Pucci, J. Lechner-Scott, R. Bergamaschi, C. Oreja-Guevara, G. Iuliano, V. Van Pesch, F. Granella, C. Ramo-Tello, D. Spitaleri, T. Petersen, M. Slee, F. Verheul, R. Ampapa, M. P. Amato, P. McCombe, S. Vucic, J. L. Sánchez Menoyo, E. Cristiano, M. H. Barnett, S. Hodgkinson, J. Olascoaga, M. L. Saladino, O. Gray, C. Shaw,

F. Moore, H. Butzkueven, T. Kalincik, on behalf of the MSBase Study Group, Defining secondary progressive multiple sclerosis, Brain 139 (9) (2016) 2395–2405. doi:10.1093/brain/aww173.

[22] L. Kappos, H. Butzkueven, H. Wiendl, T. Spelman, F. Pellegrini, Y. Chen, Q. Dong, H. Koendgen, S. Belachew, M. Trojano, On Behalf of the Tysabri® Observational Program (TOP) Investigators, Greater sensitivity to multiple sclerosis disability worsening and progression events using a roving versus a fixed reference value in a prospective cohort study, Mult Scler J 24 (7) (2018) 963–973. doi:10.1177/1352458517709619.

[23] H. Butzkueven, J. Chapman, E. Cristiano, F. Grand'Maison, M. Hoffmann, G. Izquierdo, D. Jolley, L. Kappos, T. Leist, D. Pöhlau, V. Rivera, M. Trojano, F. Verheul, J.-P. Malkowski, MSBase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis, Mult Scler J 12 (6) (2006) 769–774. doi:10.1177/1352458506070775.

[24] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, M. S. Lauer, Random survival forests, Ann Appl Stat 2 (2008) 841–860. doi:10.1214/08-AOAS169.

[25] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[26] T. Fawcett, An introduction to ROC analysis, Pattern Recogn Lett 27 (8) (2006) 861–874. doi:10.1016/j.patrec.2005.10.010.

[27] S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, Adv Neur In (2017) 4766–4775.
URL https://dl.acm.org/doi/10.5555/3295222.3295230#

[28] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles, arXiv (2019). doi:10.48550/arXiv.1802.03888.

[29] K. Dedja, F. K. Nakano, K. Pliakos, C. Vens, Bellatrex: Building explanations through a locally accurate rule extractor, IEEE Access 11 (2023) 41348–41367. doi:10.1109/ACCESS.2023.3268866.

[30] R. Martin, M. Sospedra, M. Rosito, B. Engelhardt, Current multiple sclerosis treatments have improved our understanding of MS autoimmune pathogenesis, Eur J Immunol 46 (9) (2016) 2078–2090. doi:10.1002/eji.201646485.

[31] B. A. Cree, D. L. Arnold, J. Chataway, T. Chitnis, R. J. Fox, A. Pozo Ramajo, N. Murphy, H. Lassmann, Secondary Progressive Multiple Sclerosis: New Insights, Neurology 97 (8) (2021) 378–388. doi:10.1212/WNL.0000000000012323.

[32] Y. Yang, M. Wang, L. Xu, M. Zhong, Y. Wang, M. Luan, X. Li, X. Zheng, Cerebellar and/or Brainstem Lesions Indicate Poor Prognosis in Multiple Sclerosis: A Systematic Review, Front Neurol 13 (2022) 874388. doi:10.3389/fneur.2022.874388.

[33] J. Hobart, J. Freeman, A. Thompson, Kurtzke scales revisited: the application of psychometric methods to clinical intuition, Brain 123 (5) (2000) 1027–1040. doi:10.1093/brain/123.5.1027.

[34] S. Pölsterl, N. Navab, A. Katouzian, Fast training of support vector machines for survival analysis, in: A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, C. Soares (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2015, pp. 243–259.