Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression

Edward De Brouwer[a,*], Thijs Becker[b,*], Yves Moreau[a,*], Eva Kubala Havrdova[d], Maria Trojano[e], Sara Eichau[f], Serkan Ozakbas[g], Marco Onofrj[h], Pierre Grammond[i], Jens Kuhle[j], Ludwig Kappos[j], Patrizia Sola[k], Elisabetta Cartechini[l], Jeannette Lechner-Scott[m], Raed Alroughani[n], Oliver Gerlach[o], Tomas Kalincik[p,q], Franco Granella[r], Francois Grand'Maison[s], Roberto Bergamaschi[t], Maria José Sá[u], Bart Van Wijmeersch[v], Aysun Soysal[w], Jose Luis Sanchez-Menoyo[x], Claudio Solaro[y], Cavit Boz[z], Gerardo Iuliano[aa], Katherine Buzzard[ab], Eduardo Aguera-Morales[ac], Murat Terzi[ad], Tamara Castillo Trivio[ae], Daniele Spitaleri[af], Vincent Van Pesch[ag], Vahid Shaygannejad[ah], Fraser Moore[ai], Celia Oreja-Guevara[aj], Davide Maimone[ak], Riadh Gouider[al], Tunde Csepany[am], Cristina Ramo-Tello[an], Liesbet Peeters[c,b,*]

[a] *ESAT-STADIUS, KU Leuven, 3001 Leuven, Belgium*
[b] *I-Biostat, Data Science Institute, Hasselt University, Diepenbeek, Belgium*
[c] *Department of Immunology, Biomedical Research Institute, Hasselt University, Diepenbeek, 3590, Belgium*
[d] *Charles University in Prague and General University Hospital, Prague, Czech*
[e] *Department of Basic Medical Sciences, Neuroscience and Sense Organs, University of Bari, Bari, Italy*
[f] *Hospital Universitario Virgen Macarena, Sevilla, Spain*
[g] *Dokuz Eylul University, Konak/Izmir, Turkey*
[h] *University G. d'Annunzio, Chieti, Italy*
[i] *CISSS Chaudire-Appalache, Levis, Canada*
[j] *Neurologic Clinic and Policlinic, MS Center and Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel, University of Basel, Basel, Switzerland*
[k] *Azienda Ospedaliera Universitaria, Modena, Italy*
[l] *Azienda Sanitaria Unica Regionale Marche - AV3, Macerata, Italy*
[m] *University Newcastle, Newcastle, Australia*
[n] *Amiri Hospital, Sharq, Kuwait*

[*] Corresponding authors

Email addresses: edward.debrouwer@esat.kuleuven.be (Edward De Brouwer), thijs.becker@uhasselt.be (Thijs Becker), moreau@esat.kuleuven.be (Yves Moreau), liesbet.peeters@uhasselt.be (Liesbet Peeters)

Preprint submitted to Computer Methods and Programs in Biomedicine September 28, 2021

[o] Zuyderland Ziekenhuis, Sittard, Netherlands
[p] Melbourne MS Centre, Department of Neurology, Royal Melbourne Hospital, Melbourne, Australia
[q] CORe, Department of Medicine, University of Melbourne, Melbourne, Australia
[r] University of Parma, Parma, Italy
[s] Neuro Rive-Sud, Quebec, Canada
[t] IRCCS Mondino Foundation, Pavia, Italy
[u] Department of Neurology, Centro Hospitalar Universitario de São João and University Fernando Pessoa, Porto, Portugal
[v] Rehabilitation and MS-Centre Overpelt and Hasselt University, Hasselt, Belgium
[w] Bakirkoy Education and Research Hospital for Psychiatric and Neurological Diseases, Istanbul, Turkey
[x] Hospital de Galdakao-Usansolo, Galdakao, Spain
[y] Dept of Rehabilitation mons L Novarese Hospital, Moncrivello, Italy
[z] KTU Medical Faculty Farabi Hospital, Trabzon, Turkey
[aa] previously at Ospedali Riuniti di Salerno, Salerno, Italy
[ab] Box Hill Hospital, Melbourne, Australia
[ac] University Hospital Reina Sofia, Cordoba, Spain
[ad] 19 Mayis University, Samsun, Turkey
[ae] Hospital Universitario Donostia, San Sebastain, Spain
[af] Azienda Ospedaliera di Rilievo Nazionale San Giuseppe Moscati Avellino, Avellino,Italy
[ag] Cliniques Universitaires Saint-Luc, Brussels, Belgium
[ah] Isfahan Neurosciences Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
[ai] Jewish General Hospital, Montreal, Canada
[aj] Hospital Clinico San Carlos, Madrid, Spain
[ak] Garibaldi Hospital, Catania, Italy
[al] Razi Hospital, Manouba, Tunisia
[am] University of Debrecen, Debrecen, Hungary
[an] Hospital Germans Trias i Pujol, Badalona, Spain

Abstract

Background and Objectives.

Research in Multiple Sclerosis (MS) has recently focused on extracting knowledge from real-world clinical data sources. This type of data is more abundant than data produced during clinical trials and potentially more informative about real-world clinical practice. However, this comes at the cost of less curated and controlled data sets. In this work we aim to predict disability progression by optimally extracting information from longitudinal patient

data in the real-world setting, with a special focus on the sporadic sampling problem.
Methods
We use machine learning methods suited for patient trajectories modeling, such as recurrent neural networks and tensor factorization. A subset of 6,682 patients from the MSBase registry is used.
Results
We can predict disability progression of patients in a two-year horizon with an ROC-AUC of 0.67, which represents a 11% decrease in the ranking pair error (1-AUC) and a compared to reference methods using static clinical features.
Conclusion
Compared to the models available in the literature, this work uses the most complete patient history for MS disease progression prediction and represents a step forward towards AI-assisted precision medicine in MS.

Keywords: Multiple Sclerosis, Machine Learning, Longitudinal data, Recurrent neural networks, Electronic health records, Disability progression, Real-world data

---

1. Introduction

Multiple Sclerosis (MS) is a chronic autoimmune disease characterized by heterogeneous progression across patients [1, 2]. This heterogeneity led to the clinical classification of different disease stages [3, 4, 5] with patients typically starting in the relapsing remitting (RR) phase, which can later progress to the secondary progressive (SP) phase. Clinical practice is aimed at keeping disability progression under control [1]. This led to the development of statistical methods to accurately predict the conversion from the relapsing remitting to the secondary progressive stage [6, 7]. However, to achieve more useful predictions, we would like to predict disease progression in a more detailed manner, for example using the Expanded Disability Status Scale (EDSS) [8]. The EDSS is a score designed by clinicians to quantitatively assess patient disability with improved consistency and decreased subjectivity. This paper aims at predicting disability progression on the EDSS using longitudinal clinical patient data. This longitudinal data, referred to as 'patient trajectories' here, consists of the medical follow-up of patients over time along with the most important predictors such as current and past disability

progression, past relapses, and most importantly current EDSS. Compared to previous approaches, which used mainly static information [9, 10], using the detailed clinical history of each patient is expected to increase predictive power [11, 12, 13].

One reason for not considering full patient trajectories in the past resides in the lack of data sets containing a large amount of patient-level longitudinal clinical data. Fortunately, advances in clinical practice and clinical data acquisition standards now facilitate the collection of large amounts of longitudinal data, both in terms of number of patients, but also in the number of clinical variables collected on a systematic basis. The MS community is particularly prolific in this regard with multiple international consortia, such as the MS Data Alliance [14, 15], MSBase [16, 17], Multiple MS, or Big MS Network, as well as large registries, such as the Danish, Swedish and Italian registries [18, 19, 20].

However, complex longitudinal clinical data poses challenges for modeling. It is high dimensional, consists of different data types and is sparsely measured at a non-constant sampling rate. The non-constant sampling occurs because observations are only recorded at medical visits, which can be days, months, or even years apart. Clinicians may not perform all available tests at each visit. For instance, the number of hyperintense cerebral lesions on MRI are usually not available at each clinical visit. Suitable machine learning methods should therefore need to be able to optimally extract relevant information from this type of data. Common strategies for dealing with these challenges include imputation and time binning, which lead to loss of information and thus lower performance for the predictive task of interest.

In this work, we employ several models from the machine learning literature that can deal with sporadic time series and investigate their ability to predict disability progression of MS patients using their clinical trajectories. We study several model classes: Bayesian probabilistic tensor factorization (BPTF) [21], continuous-time recurrent neural networks (RNN) [22, 23], and time-aware recurrent neural networks [24]. These models are trained on the task of predicting disability progression of individual patients over a 2-year horizon, achieving an ROC-AUC of 0.67. We used one of the largest available MS registries, MSBase, to train and validate our models. To the best of our knowledge, this work uses the most complete patient history for MS disease progression prediction.

The structure of this paper is as follows. Section 2 presents related work that uses real-world patient trajectories, with an emphasis on MS. Sections

3.1 and 3.2 provide a detailed description of the task and of the patient cohort, respectively. Section 3.3 describes the different models we propose, as well as the baselines we compare against. Sections 4 and 5 present the results of the methods we considered, their interpretations, and a vision for future work.

2. Related work

Many recent publications have used statistical models and machine learning to distil new knowledge from MS real-world clinical data [11, 25, 16]. Among them, some have developed methods using the longitudinal clinical history of the patients to predict or classify the disease course (more specifically the conversion from RRMS to SPMS) [6, 7, 26, 27, 28]. With a different focus, Signori et al. [29] used patients disability trajectories to uncover patient subgroups using latent class mixed models, and showed that those groups had different probabilities of reaching an EDSS of 6. Tacchella et al. [30] investigated the improvement of performance in prediction of MS course provided by a synergy of machine learning models and clinicians.

In contrast, our work aims at predicting disability progression, which is more specific given that patients with declining neurological capacity can remain in the same disease course category. There has been research focused on the prediction of the disability progression of MS patients, most of them using static features, and thus not considering the evolution of the patient over time. Among those, Tousignant et al. [9] used convolutional neural networks to predict prognosis from MRI scans from a single visit, achieving an ROC-AUC of around 0.70. Law et al. [10] proposed a decision tree approach based on static physiological variables. Yperman et al. [31] used random forests on features engineered on evoked potential time series to predict disability progression. Yet, to the best of our knowledge, there has been no work using longitudinal machine learning models to predict disability score progression from the full clinical history of MS patients.

3. Methods

3.1. Prediction task definition

We consider the task of predicting disability progression of patients based on their previous EDSS, DMT and relapse history. More formally, we have N multiple sclerosis patients along with a matrix $X \in \mathbb{R}^{N \times d}$ of d-dimensional

static covariates. For each patient i, we also have information about his or her medical history that we represent as a matrix $Y_i \in \mathbb{R}^{D \times N_{T_i}}$ and its corresponding vector of N_{T_i} observations at times $t_i \in N_{T_i}$ where D is the number of longitudinal variables. As every observation dimension might not be observed at every observation time, we also define a mask $M_i \in \{0,1\}^{D \times N_{T_i}}$. If an observation is missing, the entry in the mask matrix and in Y_i will be set to 0. This configuration represents what we call a sporadic time series. The timing between observations varies from patient to patient: each has its own observation times t_i and some observations might be missing at each observation time as more graphically represented on Figure 1.

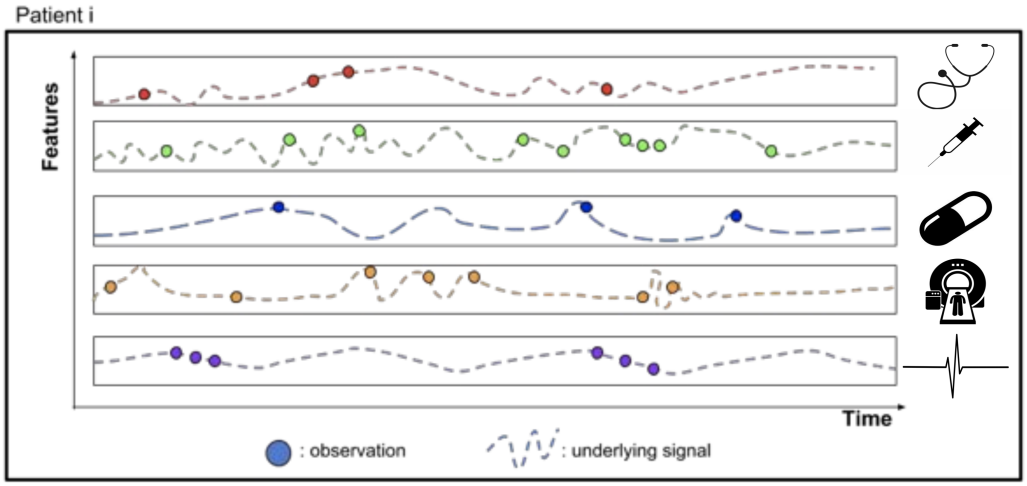


Figure 1: Illustration of sporadic time series for one patient. Dots stand for available measurements or observations while the dotted line stand for the true underlying process that would be observed in case of continuous follow-up. The sampling is very irregular in time as data is only collected during medical visits and all measurements are not sampled each time.

Our goal is to use patient covariates X_i and patient history Y_i to predict disability progression after 2 years, based on the preceding 3-year trajectory. The binary label of disability progression w after 2 years is defined as

$$w_i = \begin{cases} 1 & \text{if } \Delta_{EDSS} \geq 1.5 \ \& \ EDSS_{t_0} = 0 \\ 1 & \text{if } \Delta_{EDSS} \geq 1 \ \& \ EDSS_{t_0} \leq 5.5 \\ 1 & \text{if } \Delta_{EDSS} \geq 0.5 \ \& \ EDSS_{t_0} > 5.5 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

(2)

where $w = 1$ indicates disability progression (also referred to as worsening). This 3-strata criterion is clinically motivated in [11, 32]. It takes into account that the EDSS scale is highly nonlinear (*e.g.*, an increase of 1 point over 5.5 results in much higher impairment than for lower scores). The time indexing of each patient starts at the observation time t_0 of the baseline $EDSS_{t_0}$, as illustrated on Figure 2. So $\Delta_{EDSS} = EDSS_{t_2^*} - EDSS_{t_0}$. The variables contained in Y_i are those measured in the interval $t_i \in [-3, 0]$. In practice, it rarely happens that another observation occurs exactly 2 years after $t = 0$ so we refer to $EDSS_{t_2^*}$ as the closest observation from $t = 2$, and occurring in the interval $t \in [1, 3]$. Patients without at least one observation between $t = 1$ and $t = 3$ are therefore discarded.

To reliably assess disability progression, we use confirmed disability progression [11, 32]. We discard all EDSS measurements occurring less than 1 month after a relapse in the test period (*i.e.,* with $t > 0$) as recommended in [32]. Note that $EDSS_{t_2^*}$ can nevertheless occur less than 1 month after a relapse. Progression should be confirmed by ensuring that all EDSS measurements for at least 6 months after $EDSS_{t_2^*}$ remain above the required threshold for disability progression as defined in Equation 1. We thus require at least one confirmed EDSS measurement after $EDSS_{t_2^*}$.

Finally, we can define our task as predicting the worsening label w_i from static data X_i and historical data $Y_{i,t}$ where $t \in [-3, 0]$.

3.2. Cohort characteristics

We used the cohort of MS patients from MSBase [17], which contained at extraction time (August 2018) 55,409 unique patient records. We selected a subset of the initial cohort that complies with the following quality requirements. We first remove patients with missing or invalid diagnosis dates. This includes an invalid format or aberrant dates (dates in the future or before 1900). We remove all visit entries without EDSS value and with a date of

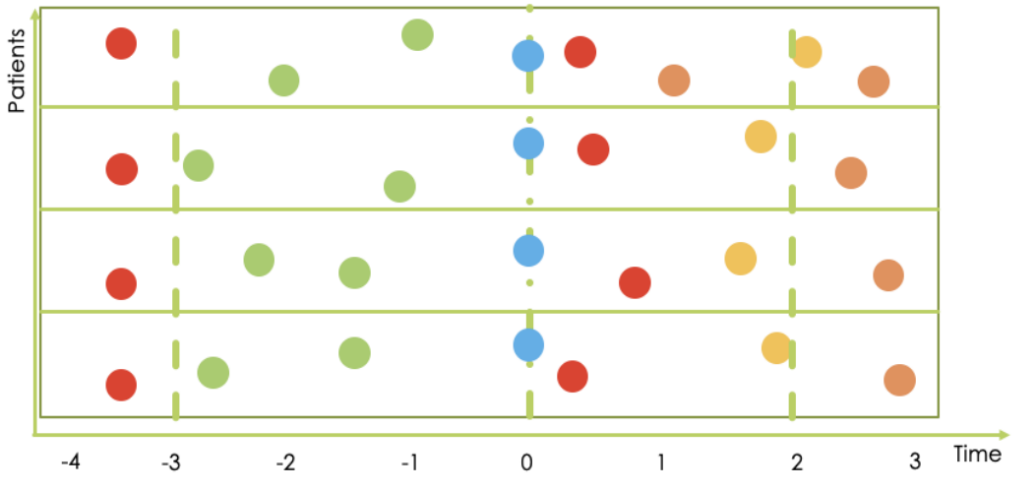


Figure 2: Graphical visualisation of the disability progression prediction task. Dots represent EDSS measurements over time. We aim at predicting disability progression at time $t = 2$ from the information available at time $t \in [-3, 0]$ (*i.e.*, we limit the EDSS progression history at 3 years back in time). The green and blue points represent the available EDSS measurements for prediction. Disability progression is defined with respect to the last observed EDSS in the observation window (blue). We define a Δ_{EDSS} as the difference between the EDSS closest to $t = 2$ (orange) and EDSS at time 0 ($t = 0$). Progression is assessed depending on the value of the last EDSS (blue) and the Δ (orange – blue) as in Equation 1. Furthermore, only confirmed progressions are considered. That is, Δ_{EDSS} that are maintained over a period of at least 6 months. Note that we discard all EDSS measurements occurring less than 1 month after a relapse in the test period ($t > 0$).

visit before the onset date. We also removed all patients with visits before 1990 or with onset date before 1990. This is done with the motivation to analyze contemporary data and building a homogeneous patient cohort in terms of standard of care.

We only selected patients with at least 6 visits in the three-year observation period (between $t = -3$ and $t = 0$) so as to have enough extra information for the trajectories.

After the cleaning procedure, we have $6,682$ patients among which $1,114$ patients progressed in disability after 2 years. Table 1 reports some summary statistics of the final cohort. In Table H.8 in Appendix H, we report a detailed description of the cohort definition process along with the number of patients discarded at each step. The preprocessing code is available at `https://github.com/edebrouwer/msbase2020`.

Attribute	Mean [C.I.]	Std	Min	Max
Total number of patients	6682	/	/	/
EDSS counts per patient ($t \in [-3, 0]$)	9.34[9.26, 9.42]	3.28	6	36
Disease duration at $t = 0$ [years]	6.88[6.78, 6.99]	4.37	3	25.12
Average EDSS per patient ($t \in [-3, 0]$)	2.38[2.34, 2.41]	1.48	0	8.5
Average EDSS per patient ($t \geq 0$)	2.67[2.62, 2.71]	1.80	0	9.1
Age at onset [years]	32.17[31.95, 32.95]	9.22	18	73.46
Female patients [% of total]	71.13%	/	/	/
CIS patients [total]	0	/	/	/
Primary Progressive patients [total]	217	/	/	/
Progressive Relapsing patients [total]	93	/	/	/
Secondary Progressive patients [total]	329	/	/	/
Relapsing-Remitting patients [total]	5722	/	/	/

Table 1: Summary statistics of the cohort of interest (mean, standard deviation, minimum value, and maximum value). For the mean, we provide the point estimate along with a 95% confidence interval (CI).

3.3. Machine Learning Methods

In this section, we define the modeling techniques used to meet the objective discussed in the previous section. We considered five models: a *static* random forest trained on only the variables that are available at $t = 0$, a *dynamic* random forest trained on engineered features representing the patient trajectory between $t = -3$ and $t = 0$, a Bayesian Probabilistic Matrix Factorization (BPMF) technique that can handle time series with missing data, a time-aware recurrent neural network, and GRU-ODE-Bayes, a continuous-time neural network model designed to deal with sporadic time series. The code for the training of the different models is available at `https://github.com/edebrouwer/msbase2020`.

3.3.1. Random forests

Random forests are popular in the statistical and machine learning community as they are robust to overfitting and more interpretable than many other machine learning methods. In particular, they have been used extensively in the MS literature [6, 26]. However, as mentioned in the introduction, those methods are not designed to take time series as input, especially if the time series is sporadic.

To overcome this problem, one usually simplifies the input data by extracting meaningful features and feeding them as a complete covariate vector to the random forest algorithm. More specifically, for each patient i, one extracts from X_i and Y_i some feature vector z_i that is fully observed. That is, each dimension of z_i can be computed for every patient. This fully observed vector can then be used along the target label w_i to train a random forest model.

The main difficulty in this type of approach is to extract informative features from the input data, which is also known as feature engineering. To highlight the information contained in the temporal medical history of the patients, we consider two sets of features: one static and one dynamic.

As their name suggests, the static feature set contains only static information about the patient, and thus nothing about temporal history, while the dynamic feature set contains information regarding the past clinical history. We now detail both feature sets.

3.3.1.1 Static feature set

In the static feature set, we ignore any past information about the patient. The features we retained in this setup are

- Gender (binary)
- Age at onset (in years)
- MS course (stage of the disease the patient is currently in [5] at time $t = 0$: RRMS, SPMS, Primary Progressive MS (PPMS), or Clinically Isolated Syndrome (CIS))
- Disease duration (years since onset at time $t = 0$)
- EDSS measured at that particular visit (at $t = 0$)
- Last used disease modifying therapies (DMT) at $t = 0$.

Note that we included MS course and disease duration, which are actually representative of the patient clinical history. However, these are non-longitudinal variables, and they are generally available to the clinicians. DMTs were grouped by degree of activity (Mild, Moderate and Highly active). For a complete description of the DMT groups used in the analysis, we refer the reader to Appendix D.

3.3.1.2 Dynamic feature set

The dynamic feature set contains the static feature set and extends it with features that are meant to reflect information from the patient's trajectory. As the history of the patient cannot be fed easily to the random forest, we have to select features that might contain relevant information in the trajectory.

On top of previously listed features, the dynamic feature set includes

- The EDSS closest to time $t = -3$, that is, the first EDSS that was measured for that patient.
- The maximum EDSS value that was reached over the observation window between $t = -3$ and $t = 0$.
- The difference between the maximum and minimum EDSS in the observation window.
- The number of visits between $t = -3$ and $t = 0$.
- The number of relapses in the observation period $t = -3$ and $t = 0$.
- The average EDSS value between $t = -3$ and $t = 0$.

Those 6 features are thought to be informative for the future course of the disease. Indeed, knowing EDSS at time $t = -3$ and $t = 0$ gives us information about the slope of progression of the disease over 3 years. The maximum EDSS and the difference between the maximum and minimum contains information of the variability of the trajectory.

3.3.2. Longitudinal Variables

On top of the static and the dynamic feature set, we consider longitudinal variables that represent the medical trajectory of the patient. Those are measured over time and consists of measurements types, along with measurements values and measurement times. In this work, we used the following available longitudinal variables:

- EDSS measurements (encoded as a value from 0 to 10).
- Relapses occurrence (encoded as a binary variable set to 1 when a relapse occurs).

- Start of a mild DMT (encoded as a binary variable set to 1 when a new mild DMT is prescribed).
- Start of a moderate DMT (encoded as a binary variable set to 1 when a new moderate DMT is prescribed).
- Start of a high DMT (encoded as a binary variable set to 1 when a new high DMT is prescribed).
- End of a mild DMT (encoded as a binary variable set to 1 when a new mild DMT is stopped).
- End of a moderate DMT (encoded as a binary variable set to 1 when a new moderate DMT is stopped).
- End of a high DMT (encoded as a binary variable set to 1 when a new high DMT is stopped).

Importantly, all variables might not be observed at every measurement time, resulting in missing values. Table 2 reports the amounts of missing values for each longitudinal variable in the cohort. The percentage is computed by taking the ratio of the number of available measurements of each type divided by the total number of measurements. Because too few patients have data about MRI measurements, we did not include information about MRI in the longitudinal measurements. Note also that in the BPTF case, because binary variables like the occurrence of a relapse or the prescription of DMT cannot be incorporated, we restrict the set of longitudinal variables to EDSS only.

Variable	**Missing %**
EDSS	12.1%
Relapse	88.3%
DMT start	91.9%
DMT end	94.5%

Table 2: Percentages of missing values for the longitudinal variables. Expressed as a fraction of the number of visits with no observation of the variable over total number of visits

3.3.3. Bayesian tensor factorization

The first method we addressed to deal with sporadically measured time series is Bayesian Probabilistic Tensor Factorization (BPTF), an extension of BPMF to tensors. In general, tensor factorization methods aim at approximating a tensor as the linear combination of r rank-1 tensors [33]. To explain why we can use Bayesian factorization techniques here, we shall first give some details about the data representation.

3.3.3.1 Data representation

We stated in the previous section that each patient history was encoded in a matrix Y_i. By reworking this data representation and stacking all patient histories together, we can have a 3-mode tensor $\mathcal{Y}$. A 3-mode tensor has 3 axes and can be best thought of as a cube. In our case, the first axis would represent the patient index, the second the measurement type (here, we have only on measurement type: EDSS), and the third would be time such that $\mathcal{Y}_{i,j,t}$ would store the measurement type j at time t for patient i. This entails two main consequences. First, the time axis is shared for all the patients, meaning that some time binning will be needed and we will lose some temporal information. Second, most of the entries in the tensor will be empty (*i.e.*, non-observed).

Binning the data in temporal bins leaves us with a trade-off. Small time steps would result in limited information loss but would make the tensor much sparser. To keep computations manageable without much temporal information loss, we chose a time bin of 30 days, which has also the advantage of being intuitive (1 month). With this binning factor, the tensor created with the data of our patients between $t = -3$ and $t = 3$ has a filling rate of 21% if we only consider EDSS.

3.3.3.2 Incorporating static features

The patient trajectories can be encoded such as to be processed by BPTF. However, static features such as gender and disease course are very important for accurate prediction. To incorporate this source of information into the model, we consider two paths. The first is BPTF with side information as presented in [21]. However, this mapping is multilinear, which restricts the possibilities for the model to extract useful, possibly nonlinear, interactions between static covariates and the worsening label. To address this issue, we

considered a second version of the model that consists of the same random forest model as described in Section 3.3.1, but where we extended the dynamic feature set with the prediction of the BPTF model at time $t = 2$: $\mathcal{Y}_{i,j=EDSS,t=3}$. We call this variant BPTF-SI-RF. More technical details for both approaches are presented in Appendix B.

3.3.4. Time-Aware Recurrent Neural Networks

Standard recurrent neural network (RNN) architectures usually require a fixed step size in between observations, an assumption that this not met in the clinical time series we aim at analyzing. Yet, one can transform sporadic data into a sequence of observation vectors along with their observation times. For an observation matrix $Y_i \in \mathbb{R}^{D \times N_{T_i}}$ and time vector $t_i \in \mathbb{R}^{N_{T_i}}$, we then build the sequence $Y_i^* \in \mathbb{R}^{(D+1) \times N_{T_i}}$ where the last row of Y_i^* consists of the observation time. We can then feed this data representation to a recurrent neural network.

In this work, we consider a Gated Recurrent Unit (GRU) variant of RNNcell [34]. GRUs are long-term memory cells and have the advantage of having fewer parameters than other models (*e.g.* Long Short Term Memory (LSTM)). We initialized the first hidden state of the GRU by feeding the static information through a multilayer perceptron (MLP) and compute the probability of worsening by feeding the last hidden state (after all observations have been processed) to another MLP. We call this model GRU-TA.

3.3.5. GRU-ODE-Bayes

The methods presented above used some artifice to deal with sporadic temporal data. The random forests use summary statistics of the trajectories, BPTF requires time binning of the time series, and time-aware RNNs consider time as if it were a feature. This has the obvious limitations of (1) losing data points (because of summarizing or of averaging in the binning case) and (2) degrading the timing accuracy of the measurements.

To more naturally accommodate the sporadic nature of the data, we use the GRU-ODE-Bayes model [23]. GRU-ODE-Bayes was recently proposed as a new method to deal with sporadic time series. It assumes a continuous latent process $h(t)$ (some hidden health status) that generates the observations $Y(t)$ and tries to approximate the dynamics of the patient as shown on Figure 1. More technical details about the approach are presented in Appendix C.

4. Results

To tune the hyperparameters, we used 5-fold cross-validation and used the exact same 5 training and validation sets over the different models for the sake of fair comparison. For the GRU-based models, we optimize the binary cross entropy. We report the average ROC-AUC and AUC-PR (precision-recall) metrics evaluated on 5 held-out test sets, as well as the standard deviation of those results. More details about training and the hyper-parameters optimization can be found in Appendix E.

Performance results are displayed in Table 3. We observe that static features only contain limited information for prediction of disability progression resulting in mediocre predictive performance (ROC-AUC of 0.63 and AUC-PR of 0.26 for the static feature set). Adding engineered temporal features improves the performances (ROC-AUC of 0.67 and AUC-PR of 0.27 for the dynamic feature set). Incorporating the full patient trajectories does not appear to significantly improve the ROC-AUC but improves the AUC-PR (ROC-AUC of 0.68 for BPTF and 0.67 for GRU-TA and GRU-ODE-Bayes, and an AUC-PR of 0.29 for all BPTF, GRU-TA and GRU-ODE-Bayes). Figure 3 presents the ROC curve and the precision-recall curves of the compared models. Importantly, these difference in performance between the static and the dynamic (or longitudinal) regimes (dynamic set and longitudinal variables) are significant. Statistical tests for significance are reported in Appendix F.

Model Type	**Model Name**	**ROC-AUC**	**AUC-PR**
	Random Model	0.5	0.16
Random Forest	Static feature set	0.63 ± 0.02	0.26 ± 0.01
Random Forest	Dynamic feature set	0.67 ± 0.01	0.27 ± 0.02
BPTF	BPTF-SI-RF	$\mathbf{0.68} \pm 0.01$	$\mathbf{0.29} \pm 0.01$
Time-aware RNN	GRU-TA	$\mathbf{0.67} \pm 0.01$	$\mathbf{0.29} \pm 0.02$
ODE-RNN	GRU-ODE-Bayes	0.66 ± 0.02	$\mathbf{0.29} \pm 0.02$

Table 3: Results for disability progression prediction with the different models. Best results are in bold. If several values are in bold, the results are not significantly different (significance assessed with pair-wise t-test).

4.1. Patient trajectory analysis

The GRU-ODE-Bayes model allows us to analyze the temporal evolution of the probability of worsening. Indeed, at each point in time, we can inte-

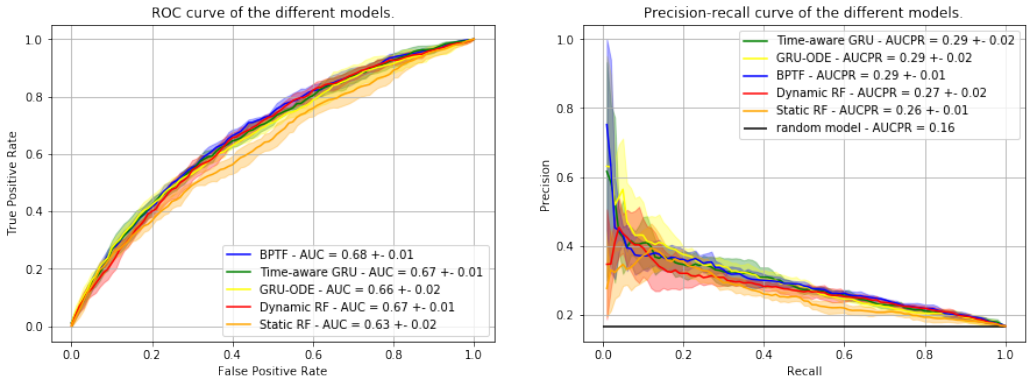


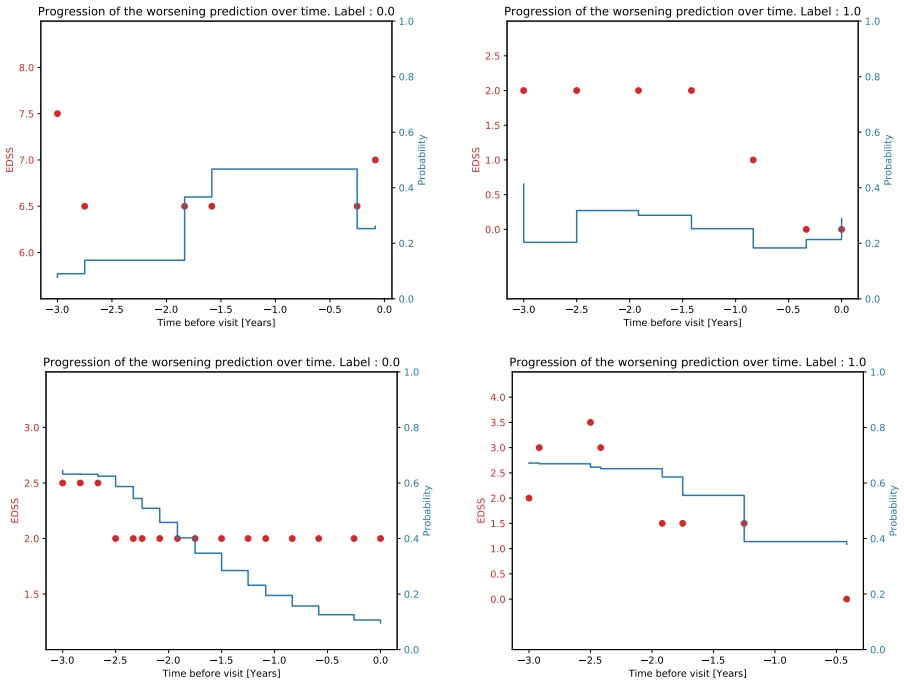
Figure 3: Receiver Operating Characteristic and Precision-Recall curves of the compared models. To avoid clutter, we do not include the curve for the BPTF-SI model.

grate the hidden process until $t = 0$ and predict the worsening label. This allows us to evaluate the impact of the sequence of EDSS measurement on the worsening prediction. Figure 4 shows an example of four randomly picked EDSS trajectories, two worsening and two non-worsening. At each point in time, we can compute the probability of worsening (from the GRU-ODE-Bayes model), would no other EDSS measurement be observed until the end of the observation period. The probability of worsening was calibrated with Platt scaling [35]. On the left column of the figure (non-worsening patients), we observe that a probability of worsening seems to decrease when a the EDSS seems to stabilize suggesting a second progression 2 years after $t = 0$ is less likely. On the right figure, for the worsening patients, we observe the same effect.

4.2. Sensitivity analysis

We performed a sensitivity analysis of the GRU-ODE-Bayes model, to assess the most predictive variables in the model. For each covariate, an importance score is determined as follows. The values of the covariate are shuffled randomly among the patients, making this covariate essentially non-predictive. The importance score is the average ROC-AUC degradation for the GRU-ODE-Bayes model, calculated from 10 repetitions (i.e., we repeat the shuffling of each covariate 10 times to improve the estimation of the average degradation effect).

Table 4 presents the importance scores for the most impactful variables. The full EDSS trajectory is the most important feature, which highlights the



(a) Non-worsening patients.

(b) Worsening patients.

Figure 4: Evolution of the predicted probability of worsening for 2 sampled patients EDSS trajectories. Predictions are performed with GRU-ODE-Bayes. The blue line represents the probability of worsening at each point in time if no further measurement would be observed. The red dots stand for the observed EDSS values for that particular patient.

importance of the EDSS history of the patient in the prognosis. Furthermore, we see that the maximum EDSS value, the previous EDSS and the average EDSS in the observation window all rank high in terms of feature importance, further strengthening the signal contained in the EDSS trajectory. Lastly, the label indiacting if patients are in the secondary progressive stage of the disease or not appears important as well for the prediction. Other features are shown to be less important in the prediction of the worsening.

Feature	**Sensitivity Score**
Full EDSS trajectory	0.115 ± 0.044
Max EDSS	0.011 ± 0.024
Previous EDSS	0.008 ± 0.018
Mean EDSS	0.008 ± 0.002
Secondary Progressive	0.008 ± 0.02
Others	≤ 0.008

Table 4: Sensitivity analysis of the features used in GRU-ODE-Bayes. Feature are presented by order of importance, together with their standard deviations.

4.3. Performance per type of MS

The cohort of patients used in this study contains patients with different types of MS at time $t = 0$. Those patient groups are known to have specific patterns in their progression. For our models to deal with these different types of patients, we feed the current MS course as a static variable. This allows the model to tailor its prediction to the targeted class of patients and also to make use of information that is shared among the different groups. In Figure 5 we report the performance of the models split by type of MS (relapsing-remitting, secondary progressive and primary progressive). The few progressive relapsing patients were merged in the more general primary progressive class. A table with the numerical values is presented in Table G.7 in Appendix G. We observe that the performance of the relapsing-remitting cohort is similar to the main cohort as it represents the major part of it. A decrease in performance is observed for the secondary progressive and primary progressive cohort, except for the Time-aware GRU model that is conserves similar performance in the secondary progressive cohort. The main cause of this decrease is likely the low number of patients in these categories. This is further reflected by the large standard deviations of the performances of these cohort. However it is important to notice the consistency with which

models incorporating clinical historical data outperform the static baselines, highlighting the importance of longitudinal data in all types of MS.

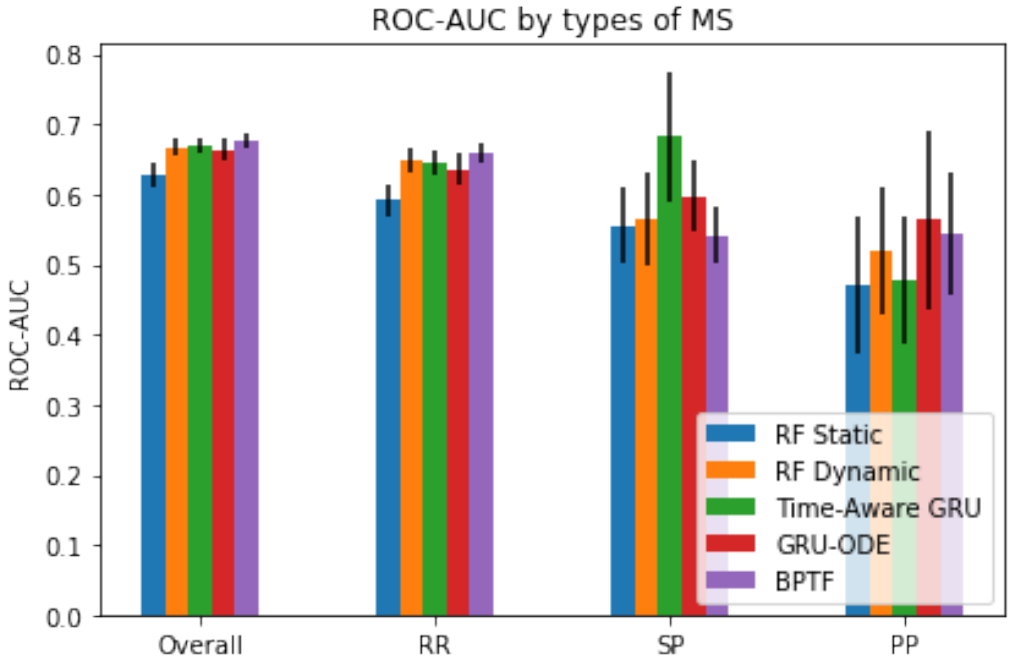


Figure 5: Area under the curve of the receiver operating characteristic (ROC-AUC) of the compared models by type of MS (relapsing-remitting, secondary progressive and primary progressive). One standard deviation of the performance of the 5 different data splits is shown as vertical black bars.

4.4. Inference Times

Prediction in medical practice can be time-critical. In particular, for a clinical MS application, it would be desirable that predictions could be performed in a reasonable amount of time, such that clinicians can take it into account and share the prognosis during a medical visit. In Table 5 we report the inference times per patient for each method. Most of them are on the order of milliseconds, except for Bayesian Tensor Factorization which is around a quarter of a second, which can be still considered instantaneous for the intended application.

Because medical application can be time-sensitive, we report below the average inference times per patient of the different models. Those number

then give an idea of the time required for performing a prediction for a single patient. Averages and standard deviations are reported in Table 5. We see that all inference times are below a second, which is compatible with online prediction in clinical practice in the MS context.

	Static	Dynamic	BPTF-SI-RF	GRU-TA	GRU-ODE
Times[ms]	0.1 ± 0.02	0.1 ± 0.05	253.3 ± 21.2	2.0 ± 0.1	4.0 ± 0.2

Table 5: Average time in milliseconds (along with standard deviations) required to perform inference for a single patient.

5. Discussion

5.1. Results interpretation and impact for clinical practice

Our results provide evidence that using the patient history results in more accurate prediction of future disability progression. Indeed, by adding simple summary features of the clinical history of the patients, we obtained an increase of performance of 0.04 points of ROC-AUC and 0.01 points of AUC-PR. This improvement was slightly more pronounced when including the full longitudinal trajectories in the modeling. In terms of ROC-AUC, BPTF (with a RF on top) and GRU-TA lead to the best performance. As BPTF does not use longitudinal information about DMT and relapses, it suggests that this information is not critical for the longitudinal models to achieve their performance. It is also not reported as significantly important for the model in the sensitivity analysis. This might be explained partially by the fact that the information about the current DMT is already given in the static feature set and that knowledge of older treatment history might not be as relevant for disability progression. Similarly, the total number of relapses is already included in the dynamic feature set. The added information about the specific timing of the relapses appears to not be very informative.

The relevance of predictive models in clinical practice hinges on their capacity to detect with high precision all the patients that will experience disability progression in the future. To correctly detect 70% of progressing patients (recall = 0.7), the static method would lead to a precision of 21%. 79% of the patients predicted as positive would then be false positives. With the full trajectory methods, the precision jumps to 24%, which represents a significant improvement that brings these models closer to being useful for

clinical practice. To put things in perspective, let us consider a hypothetical cohort of 1,000 patients with similar statistics as the MSBase one. Out of those 1,000 patients, about 160 would eventually progress after 2 years. The static method could predict around 112 of those (70%) and would wrongly detect approximately 421 patients as positives, which is more than two fifths of the full cohort. The full trajectories method, on the other hand, would detect the same number of positives, but with a lower number of false positives: only about 354, or a difference of 67 patients on a 1,000 cohort. This increase in precision leads to a more efficient clinical care as the limited resources of neurologists can be focused on a smaller and more specific subset of patients requiring special attention.

To assess quantitatively the information content in the longitudinal trajectories, we performed a sensitivity analysis with random permutations. It appeared that the EDSS trajectory was crucial for prognosis. This is in line with clinical experience [10, 36]. The three other most important features confirm this finding as they are all related to the patient clinical history (full trajectory, max EDSS, EDSS difference). However, our predictions are averaged over the whole patient cohorts and we did not assess the impact on performance for a subset of patients at an advanced stage of disability.

Finally, we studied the performance of these models per type of MS and showed that the methods using data from the clinical history of the patients consistently outperform the static baselines for all types of MS. This size of this effect was reduced for primary progressive patients however. Absolute performance for primary progressive and primary relapsing patients were also lower than for the main cohort. These two types are also the least represented with only 3.2% and 1.4% of the data respectively.

5.2. Interpretability of the models

Despite the quantitative evidence that taking past clinical trajectories into account for prognosis is beneficial, it is not yet clear which specific patterns in the trajectory are characteristics of future progression. From the example trajectories we provided in the results section, two main trends tend to appear. First, when a patient with initially stable EDSS experience some recovery (EDSS decreases), our model predicts a higher probability of worsening, suggesting the recovery is most probably temporary and the patient will progress over time. Second, a patient with initially stable EDSS experiencing a progression during the observation window has a lower probability of future worsening, suggesting a patient having worsened significantly over

the observation window is less likely to progress again afterwards. This effect could be associated with a form of regression to the mean which would be picked up by the models. However, as the mean EDSS and the last EDSS are present in the dynamic features set, the longitudinal model appear to capture a signal which is more complex than a simple reversion to the mean. Those interpretations are still qualitative and speculative and the design of dedicated methods to uncover specific patterns for prognosis of progression is left for future work.

5.3. Sources of bias

The cohort of patients we used in this study is one of the largest available cohorts in the MS research community. However, not all countries are represented in the registry which represents a potential source of bias as clinical practice standards can differ among different countries.

A selection bias is also possible, as MS centers interested in sharing their data in initiatives like MSBase might not be fully representative of the general population. Yet, because the cohort spans multiple countries on different continents, we expect the subset of patients considered to be more representative of the general population than data from a single national registry. For implementation in clinical practice, an external validation cohort would be beneficial. However, we did not have access to such an external cohort in this study.

The objective of this work being to assess the importance of using longitudinal information for disability progression prognosis, we naturally select patients having some longitudinal information available. We therefore selected a minimum of 6 visits to provide a clear evaluation of the impact of longitudinal clinical data in MS. This corresponds to a cohort that has an average of 2 visits a year, which is a realistic visit frequency for patients under follow-up in MS centers. Due to some inherent limitations of the compared models, patients lacking basic information such as gender, age or MS course were also discarded. In practice, however, this means that the models presented in this work are not directly applicable to patients with no or limited medical history. We are leaving the extension of these models to a broader class of patients for future work. We hope that our work leads to the inclusion of full trajectories in future modeling efforts in MS. Our results underline the importance of longitudinal data collection in clinical practice.

6. Conclusion

In this study, we showed that including a more complete disability history of the patient in the statistical modeling improves the predictive performance of disability progression in MS. We considered several methodologies to incorporate those sporadic trajectories for the prediction of disability worsening of MS patients, eventually achieving state-of-the-art performance and showing quantitatively the impact of including the full longitudinal trajectories in the modeling. This analysis confirms the importance of using longitudinal data to achieve AI-assisted precision medicine in MS. Indeed, we demonstrated an improvement of 3 points of precision and 10 points of specificity at a recall of 0.70, which translates into a more efficient stratification of patients to provide patients with optimal medical attention.

The evidence we provided in this paper suggests that more systematic collection of longitudinal patient data would be beneficial to patient followup, and that more accurate patient stratification and prognosis, based on the whole patient clinical history, will result in better and more patient-specific care in MS. One extra milestone towards this goal is to assess the efficacy of drugs more accurately than before (*i.e.,* using the full trajectory to detect treatment response to a newly administered DMT). Cutting significantly the amount of time required for evaluating the effectiveness of a given treatment would result in lower disability progression during the optimal treatment search period. As treatment information (drug prescriptions) is available in the MSBase registry, we leave this temporal analysis of drug efficacy for future work.

Acknowledgements and declarations

We would like to thank all patients and their caregivers who have participated in this study and who have contributed data to the MSBase cohort. The list of MSBase study group contributors are provided in Appendix A.

Yves Moreau is funded by Research Council KU Leuven: C14/18/092 SymBioSys3; CELSA-HIDUCTION CELSA/17/032 Flemish Government:IWT: Exaptation, PhD grants FWO 06260 (Iterative and multi-level methods for Bayesian multirelational factorization with features). This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. EU: "MELLODDY" This project has received funding from the Innovative Medicines Initiative 2 Joint

Undertaking under grant agreement No 831472. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. Edward De Brouwer is funded by a FWO-SB grant.

We received ethical approval for this study from the medical ethics committee of the University of Hasselt, number CME2019/059.

The authors declare the following conflicts of interest.

- 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; has been supported by the Czech Ministry of Educaon research project PROGRES Q27/LF1.

- Maria Trojano received speaker honoraria from Biogen-Idec, Bayer-Schering, Sanofi Aventis, Merck, Teva , Novartis and Almirall; has received research grants for her Institution from Biogen-Idec, Merck, and Novartis.

- Francesco Patti received speaker honoraria and advisory board fees from Almirall, Bayer, Biogen, Celgene, Merck, Novartis, Roche, Sanofi-Genzyme and TEVA. He received research funding from Biogen, Merck, FISM (Fondazione Italiana Sclerosi Multipla), Reload Onlus Association and University of Catania.

- Guillermo Izquierdo received speaking honoraria from Biogen, Novartis, Sanofi, Merck, Roche, Almirall and Teva.

- Sara Eichau received speaker honoraria and consultant fees from Biogen Idec, Novartis, Merck, Bayer, Sanofi Genzyme, Roche and Teva.

- Marc Girard received consulting fees from Teva Canada Innovation, Biogen, Novartis and Genzyme Sanofi; lecture payments from Teva Canada Innovation, Novartis and EMD . He has also received a research grant from Canadian Institutes of Health Research.

- Pierre Duquette served on editorial boards and has been supported to attend meetings by EMD, Biogen, Novartis, Genzyme, and TEVA Neuroscience. He holds grants from the CIHR and the MS Society of Canada and has received funding for investigator-initiated trials from Biogen, Novartis, and Genzyme.

- Pierre Grammond is a Merck, Novartis, Teva-neuroscience, Biogen and Genzyme advisory board member, consultant for Merck , received payments for lectures by Merck , Teva-Neuroscience and Canadian Multiple sclerosis society, and received grants for travel from Teva-Neuroscience and Novartis.

- Patrizia Sola served on scientific advisory boards for Biogen Idec and TEVA, she has received funding for travel and speaker honoraria from Biogen Idec, Merck , Teva, Sanofi Genzyme, Novartis and Bayer and research grants for her Institution from Bayer, Biogen, Merck , Novartis, Sanofi, Teva.

- Jeannette Lechner-Scott travel compensation from Novartis, Biogen, Roche and Merck. Her institution receives the honoraria for talks and advisory board commitment as well as research grants from Biogen, Merck, Roche, TEVA and Novartis.

- Raed Alroughani received honoraria as a speaker and for serving on scientific advisory boards from Bayer, Biogen, GSK, Merck, Novartis, Roche and Sanofi-Genzyme.

- Franco Granella received an institutional research grant from Biogen and Sanofi Genzyme, served on scientific advisory boards for Biogen, Novartis, Merck, Sanofi Genzyme and Roche, received funding for travel and speaker honoraria from Biogen, Merck, and Sanofi-Aventis.

- Francois Grand'Maison received honoraria or research funding from Biogen, Genzyme, Novartis, Teva Neurosciences, Mitsubishi and ONO Pharmaceuticals.

- Roberto Bergamaschi received speaker honoraria from Bayer Schering, Biogen, Genzyme, Merck , Novartis, Sanofi-Aventis, Teva; research grants from Bayer Schering, Biogen, Merck , Novartis, Sanofi-Aventis, Teva; congress and travel/accommodation expense compensations by Almirall, Bayer Schering, Biogen, Genzyme, Merck , Novartis, Sanofi-Aventis, Teva.

- Bart Van Wijmeersch received research and rravel grants, honoraria for MS-Expert advisor and Speaker fees from Bayer-Schering, Biogen, Sanofi Genzyme, Merck, Novartis, Roche and Teva.

- Jose Luis Sanchez-Menoyo accepted travel compensation from Novartis and Biogen, speaking honoraria from Biogen, Novartis, Sanofi, Merck , Almirall, Bayer and Teva and has participated in a clinical trial by Biogen.
- 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.
- Cavit Boz received conference travel support from Biogen, Novartis, Bayer-Schering, Merck and Teva; has participated in clinical trials by Sanofi Aventis, Roche and Novartis.
- Gerardo Iuliano had travel/accommodations/meeting expenses funded by Bayer Schering, Biogen, Merck , Novartis, Sanofi Aventis, and Teva.
- Katherine Buzzard received honoraria and consulting fees from Biogen, Teva, Novartis, Genzyme-Sanofi, Roche, Merck, CSL and Grifols.
- Murat Terzi received travel grants from Novartis, Bayer-Schering, Merck and Teva; has participated in clinical trials by Sanofi Aventis, Roche and Novartis.
- Tamara Castillo Triviño received speaking/consulting fees and/or travel funding from Bayer, Biogen, Merck, Novartis, Roche, Sanofi-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.
- Fraser Moore participated in clinical trials sponsored by EMD Serono and Novartis.
- Celia Oreja-Guevara received honoraria as consultant on scientific advisory boards from Biogen, Celgene, Merck, Novartis, Roche, Sanofi-Genzyme and TEVA.
- Davide Maimone received speaker honoraria for Advisory Board and travel grants from Almirall, Biogen, Merck, Novartis, Roche, Sanofi-Genzyme, and Teva.

- Tunde Csepany received speaker honoraria/ conference travel support from Bayer Schering, Biogen, Merck , Novartis, Roche, Sanofi-Aventis and Teva.

- Cristina Ramo-Tello received research funding, compensation for travel or speaker honoraria from Biogen, Novartis, Genzyme and Almirall.

References

[1] H. L. Weiner, The challenge of multiple sclerosis: how do we cure a chronic heterogeneous disease?, Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 65 (2009) 239–248.

[2] H. F. McFarland, R. Martin, Multiple sclerosis: a complicated picture of autoimmunity, Nature immunology 8 (2007) 913–919.

[3] D. H. Miller, S. M. Leary, Primary-progressive multiple sclerosis, The Lancet Neurology 6 (2007) 903–912.

[4] C. Confavreux, S. Vukusic, T. Moreau, P. Adeleine, Relapses and progression of disability in multiple sclerosis, New England Journal of Medicine 343 (2000) 1430–1438.

[5] F. D. Lublin, S. C. Reingold, et al., Defining the clinical course of multiple sclerosis: results of an international survey, Neurology 46 (1996) 907–911.

[6] A. Ion-Mărgineanu, G. Kocevar, C. Stamile, D. M. Sima, F. Durand-Dubief, S. Van Huffel, D. Sappey-Marinier, Machine learning approach for classifying multiple sclerosis courses by combining clinical data with lesion loads and magnetic resonance metabolic features, Frontiers in neuroscience 11 (2017) 398.

[7] Y. Zhao, B. C. Healy, D. Rotstein, C. R. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, T. Chitnis, Exploration of machine learning techniques in predicting multiple sclerosis disease course, PLoS One 12 (2017).

[8] J. F. Kurtzke, Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (edss), Neurology 33 (1983) 1444–1444.

[9] A. Tousignant, P. Lemaître, D. Precup, D. L. Arnold, T. Arbel, Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data, in: International Conference on Medical Imaging with Deep Learning, 2019, pp. 483–492.

[10] M. T. Law, A. L. Traboulsee, D. K. Li, R. L. Carruthers, M. S. Freedman, S. H. Kolind, R. Tam, Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression, Multiple Sclerosis Journal–Experimental, Translational and Clinical 5 (2019) 2055217319885983.

[11] T. Kalincik, A. Manouchehrinia, L. Sobisek, V. Jokubaitis, T. Spelman, D. Horakova, E. Havrdova, M. Trojano, G. Izquierdo, A. Lugaresi, et al., Towards personalized therapy for multiple sclerosis: prediction of individual treatment response, Brain 140 (2017) 2426–2443.

[12] T. Ziemssen, R. Kern, K. Thomas, Multiple sclerosis: clinical profiling and data collection as prerequisite for personalized medicine approach, BMC neurology 16 (2016) 124.

[13] H. Vrenken, M. Jenkinson, M. Horsfield, M. Battaglini, R. Van Schijndel, E. Rostrup, J. Geurts, E. Fisher, A. Zijdenbos, J. Ashburner, et al., Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis, Journal of neurology 260 (2013) 2458–2471.

[14] L. M. Peeters, T. Parciak, C. Walton, L. Geys, Y. Moreau, E. De Brouwer, D. Raimondi, A. Pirmani, T. Kalincik, G. Edan, et al., Covid-19 in people with multiple sclerosis: A global data sharing initiative, Multiple Sclerosis Journal 26 (2020) 1157–1162.

[15] L. M. Peeters, T. Parciak, D. Kalra, Y. Moreau, E. Kasilingam, P. van Galen, C. Thalheim, B. Uitdehaag, P. Vermersch, N. Hellings, et al., Multiple sclerosis data alliance–a global multi-stakeholder collaboration to scale-up real world data research, Multiple sclerosis and related disorders 47 (2021) 102634.

[16] 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, Nature Reviews Neurology 13 (2017) 105.

[17] H. Butzkueven, J. Chapman, E. Cristiano, F. Grand'Maison, M. Hoffmann, G. Izquierdo, D. Jolley, L. Kappos, T. Leist, D. Poehlau, et al., Msbase: an international, online registry and platform for collaborative outcomes research in multiple sclerosis, Multiple Sclerosis Journal 12 (2006) 769–774.

[18] N. Koch-Henriksen, The danish multiple sclerosis registry: a 50-year follow-up, Multiple Sclerosis Journal 5 (1999) 293–296.

[19] J. Hillert, L. Stawiarz, The swedish ms registry–clinical support tool and scientific resource, Acta Neurologica Scandinavica 132 (2015) 11–19.

[20] M. Trojano, R. Bergamaschi, M. P. Amato, G. Comi, A. Ghezzi, V. Lepore, M. G. Marrosu, P. Mosconi, F. Patti, M. Ponzio, et al., The italian multiple sclerosis register, Neurological Sciences 40 (2019) 155–165.

[21] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, Y. Moreau, Macau: Scalable bayesian factorization with high-dimensional side information using mcmc, in: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2017, pp. 1–6.

[22] T. Q. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, Neural ordinary differential equations, in: Advances in Neural Information Processing Systems, 2018, 2018.

[23] E. De Brouwer, J. Simm, A. Arany, Y. Moreau, Gru-ode-bayes: Continuous modeling of sporadically-observed time series, in: Advances in Neural Information Processing Systems, 2019, pp. 7377–7388.

[24] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, J. Zhou, Patient subtyping via time-aware lstm networks, in: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 65–74.

[25] J. A. Cohen, M. Trojano, E. M. Mowry, B. M. Uitdehaag, S. C. Reingold, R. A. Marrie, Leveraging real-world data to investigate multiple sclerosis

disease behavior, prognosis, and treatment, Multiple Sclerosis Journal 26 (2020) 23–37.

[26] R. Seccia, D. Gammelli, F. Dominici, S. Romano, A. C. Landi, M. Salvetti, A. Tacchella, A. Zaccaria, A. Crisanti, F. Grassi, et al., Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis, PloS one 15 (2020) e0230219.

[27] M. F. Pinto, H. Oliveira, S. Batista, L. Cruz, M. Pinto, I. Correia, P. Martins, C. Teixeira, Prediction of disease progression and outcomes in multiple sclerosis with machine learning, Scientific reports 10 (2020) 1–13.

[28] F. Pellegrini, M. Copetti, M. P. Sormani, F. Bovis, C. de Moor, T. P. Debray, B. C. Kieseier, Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling, Multiple Sclerosis Journal 26 (2020) 1828–1836.

[29] A. Signori, G. Izquierdo, A. Lugaresi, R. Hupperts, F. Grand'Maison, P. Sola, D. Horakova, E. Havrdova, A. Prat, M. Girard, et al., Long-term disability trajectories in primary progressive ms patients: A latent class growth analysis, Multiple Sclerosis Journal 24 (2018) 642–652.

[30] A. Tacchella, S. Romano, M. Ferraldeschi, M. Salvetti, A. Zaccaria, A. Crisanti, F. Grassi, Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study, F1000Research 6 (2017).

[31] J. Yperman, T. Becker, D. Valkenborg, V. Popescu, N. Hellings, B. Van Wijmeersch, L. M. Peeters, Machine learning analysis of motor evoked potential time series to predict disability progression in multiple sclerosis, BMC Neurol 20 (2020) 1–15.

[32] T. Kalincik, G. Cutter, T. Spelman, V. Jokubaitis, E. Havrdova, D. Horakova, M. Trojano, G. Izquierdo, M. Girard, P. Duquette, et al., Defining reliable disability outcomes in multiple sclerosis, Brain 138 (2015) 3287–3298.

[33] A. Mnih, R. R. Salakhutdinov, Probabilistic matrix factorization, in: Advances in neural information processing systems, 2008, pp. 1257–1264.

[34] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078 (2014).

[35] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Advances in large margin classifiers 10 (1999) 61–74.

[36] N. Lizak, A. Lugaresi, R. Alroughani, J. Lechner-Scott, M. Slee, E. Havrdova, D. Horakova, M. Trojano, G. Izquierdo, P. Duquette, et al., Highly active immunomodulatory therapy ameliorates accumulation of disability in moderately advanced and advanced multiple sclerosis, Journal of Neurology, Neurosurgery & Psychiatry 88 (2017) 196–203.

Appendix A. MSBase investigators

This study would have been possible without the help of all MSBase investigators who contributed with patients data. By order of number of contributed patients :

Eva Kubala Havrdova, Dana Horakova, Maria Trojano, Francesco Patti, Guillermo Izquierdo, Sara Eichau, Serkan Ozakbas, Marco Onofrj, Alexandre Prat, Marc Girard, Pierre Duquette, Pierre Grammond, Jens Kuhle, Ludwig Kappos, Patrizia Sola, Elisabetta Cartechini, Jeannette Lechner-Scott, Raed Alroughani, Oliver Gerlach, Tomas Kalincik, Franco Granella, Francois Grand'Maison, Roberto Bergamaschi, Maria Jose Sa, Bart Van Wijmeersch, Aysun Soysal, Ricardo Fernandez Bolaños, Jose Luis Sanchez-Menoyo, Claudio Solaro, Cavit Boz, Gerardo Iuliano, Katherine Buzzard, Olga Skibina, Julie Prevost, Eduardo Aguera-Morales, Murat Terzi, Tamara Castillo Triviño, Daniele Spitaleri, Maria Edite Rio, Vincent Van Pesch, Vahid Shaygannejad, Mark Slee, Fraser Moore, Celia Oreja-Guevara, Davide Maimone, Riadh Gouider, Tunde Csepany, Cristina Ramo-Tello, Edgardo Cristiano, Juan Ignacio Rojas, Shlomo Flechter, Maria Laura Saladino, Steve Vucic, Koen de Gans, Pamela McCombe, Radek Ampapa, Ayse Altintas, Norma Deri, Michael Barnett, Ernest Butler, Claudio Gobbi, Jose Antonio Cabrera-Gomez, Thor Petersen, Suzanne Hodgkinson, Richard Macdonell, Tatjana Petkovska-Boskova, Maria Pia Amato, Jose Andres Dominguez, Jabir Alkhaboori, Carlos Vrech, Guy Laureys, Gabor Lovas, Allan Kermode, Cameron Shaw, Anneke van der Walt, Helmut Butzkueven, Nikolaos Grigoriadis, Piroska Imre, Talal Al-Harbi, Neil Shuey, Angel Perez Sempere, Orla Gray, Magdolna Simo, Eniko Dobos, Cecilia Rajda, Bhim Singhal, Recai Turkoglu, Clara Chisari, Emanuele D'Amico, Lo Fermo Salvatore, Giovanna De Luca, Valeria Di Tommaso, Daniela Travaglini, Erika Pietrolongo, Maria di Ioia, Deborah Farina, Luca Mancinelli, Catherine Larochelle, Francesca Vitetta, Anna Maria Simone, Matteo Diamanti, Mark Marriott, Trevor Kilpatrick, John King, Katherine Buzzard, Ai-Lan Nguyen, Chris Dwyer, Mastura Monif, Izanne Roos, Lisa Taylor, Josephine Baker, Erica Curti, Elena Tsantes, Javier Olascoaga, Juan Ingacio Rojas and Freek Verheul.

Appendix B. BPTF additional details

Appendix B.1. Model details

The Bayesian tensor factorization setting posits a specific multilinear generative model for the data. First, some latent matrices are generated

from some prior for each of the modes of the tensor. Here, we have three : $U \in \mathbb{R}^{N \times r}$ for the patients axis, $V \in \mathbb{R}^{D \times r}$ for the measurements types axis and $W \in \mathbb{R}^{T_b \times r}$ for the time dimension where T_b is the number of time bins between $t = -3$ and $t = 3$. We consider the following generative process :

$$\begin{aligned} U_{i,:} &\sim \mathcal{N}(\mu_a, \Sigma_a) \text{ for } i \text{ in } 1...N \\ V_{j,:} &\sim \mathcal{N}(\mu_b, \Sigma_b) \text{ for } j \text{ in } 1...D \\ W_{t,:} &\sim \mathcal{N}(\mu_c, \Sigma_c) \text{ for } t \text{ in } 1...T_b \\ \mathcal{Y}_{i,j,t} &\sim \mathcal{N}(\sum_{k=1}^{K} U_{i,k} V_{j,k} W_{t,k}, \alpha^{-1}), \end{aligned} \tag{B.1}$$

where μ and Σ stand for the means and covariance matrices of the prior distributions. The inference then consists in identifying the posterior probability distributions of the latent matrices U, V and W conditionally on the observed values of $\mathcal{Y}$:

$$\mathbb{P}(U, V, W \mid \mathcal{Y}_{i,j,t} \text{ for each observed } (i,j,t) \text{ tuple})$$

This inference is performed using Markov Chain Monte Carlo (MCMC) techniques and more specifically Gibbs sampling as shown in [21]. Once we have computed the posterior probability of the latent matrices U, V and W, we can compute the posterior distribution of unseen samples, such as future EDDS values of patients. This can be computed using

$$\mathbb{P}(Y_{i,j,t^*} = y \mid \mathcal{Y}_{i,j,t} \text{ for each observed } (i,j,t) \text{ tuple }) =$$
$$\iiint \mathbb{P}(Y_{i,j,t^*} y \mid U, V, W) \cdot \mathbb{P}(U, V, W \mid \mathcal{Y}_{i,j,t} \text{ for each observed } (i,j,t) \text{ tuple}) dU dV dW,$$

where the first term is given by the generation model B.1 and second term is the posterior whose samples are generated by the MCMC routine.

Appendix B.2. Adding side information

To incorporate the static information we have about each patient, we use the Bayesian tensor factorization with side information framework. It consists in adding a linear mapping (a vector β from the matrix of static

covariates X to the corresponding latents). In our case, we have such information for patients only, not for the other modalities. We then update the generation process as

$$\begin{aligned} \beta &\sim \mathcal{N}(\mu_\beta, \Sigma_\beta) \\ U_{i,:} &\sim \mathcal{N}(\mu_a + \beta X_i, \Sigma_a) \text{ for } i \text{ in } 1...N \end{aligned}$$

During inference, the posterior distribution of the vector β will then have to be sampled as well. For this inference to be possible, the mapping between static covariates X and the latents has to be linear. Would the mapping be more complex, it would quickly become intractable. For this reason, in practice, we merge the prediction of the BPTF with a random forest.

Appendix C. GRU-ODE model

Appendix C.1. GRU-ODE

The GRU-ODE module parametrizes the dynamics of the latent process $h(t)$ with an Neural-ODE inspired from the classical GRU module. We use the following parametric ODE as suggested in [23]:

$$\frac{dh(t)}{dt} = (1 - z(t)) \odot (g(t) - h(t)), \tag{C.1}$$

where $\odot$ is the Hadamard product and $z(t)$ and $g(t)$ are given as in the GRU equations:

$$\begin{aligned} r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ g_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \end{aligned} \tag{C.2}$$

Appendix C.2. GRU-Bayes

GRU-Bayes module is responsible for the update of the hidden state when new measurements are observed. As data comes in in packets, we allow the hidden process to jump to a new point in hidden space where it reflects more the newly observed data point.

This update is performed by using a GRU cell that takes as input the previous hidden state and the current observation and then mimics a Bayesian update to set the hidden to a new value that matches the current observations:

$$h(t_+) = GRU(h(t_-), f(\mathbf{y}[k], m[k], h(t_-))) \tag{C.3}$$

where t_- and t_+ stand for the value of the vectors just before and after the update.

Appendix C.2.1. All together

At test time, we first compute the initial hidden state value $h(0)$ from the static covariates X with some neural network mapping $g(\cdot)$:

$$h(t = -3) = g(X).$$

We integrate the hidden process according to the GRU-ODE dynamics until the first observation (done with numerical integration). When an observation is reached, we process it with GRU-Bayes and update the hidden state. We then resume to GRU-ODE integration from the new initial point and continue until a next observation is reached. At each point in time, we can use $f_{obs}(.)$ to predict the distribution of the measurements. When we run out of observations, the predictions over time are only performed by integrating GRU-ODE until the prediction time of interest.

Appendix D. DMTs used in the analysis

We restricted the analysis to the following disease modifying therapies:

- Interferons
- Natalizumab
- Fingolimod
- Teriflunomide
- Dimethyl-Fumarate
- Glatiramer
- Alemtuzumab

- Rituximab
- Cladribine
- Ocrelizumab
- Other Siponimod
- Daclizumab

These DMTs were subsequently categorized by their efficacies. We used the following grouping :

- Mild DMT : Interferons, Teriflunomide, Glatiramer
- Moderate DMT : Fingolimod, Dimethyl-Fumarate, Cladribine, Siponimod, Daclizumab
- Strong DMT : Alemtuzumab, Rituximab, Ocrelizumab, Natalizumab, Mitoxantrone

An extra category, *No DMT*, was also introduced for patients not prescribed with a DMT at a specific time point.

Appendix E. Hyperparameter selection for the different methods

All the methods are trained using data from time $t = -3$ to $t = 0$ only such that no future information is available to the models to predict future disability prediction.

We train the models on a training set, evaluate the results of the different models hyper-parameters on a validation set and finally report the results on a left-out test set. The test set represents 15% of the whole available data. The training and validation set represent 80% and 20% of the remaining 85% of the data. In order to assess the statistical significance of our results, we generate 5 different instances of such splits, or 5 *folds*.

Each method has a different set of hyper-parameters and below we report the range of hyper-parameters that we considered for each method.

Random Forests Models

For the random forests models (both static feature set and dynamic feature set), we use a range of values for each hyper-parameter. That is, we search the optimal set of hyper-parameters in the ranges provided below :

- Number of estimators : [100,1000]
- Maximum Depth : [5,25]
- Minimum Samples Split : [2,10]

BPTF

BPTF models have only the number of latent factors as hyper-parameters as the priors of the methods have been made uninformative. We used 50 and 70 as candidates for the number of latent factors.

GRU-TA

For the time-aware GRU, we used the following hyper-parameter values. Please note that, those are pointwise values and not a range as in the random forests case. The choice of a smaller set of hyper-parameters for this model is motivated by the longer time required for training. A broader set of hyper-parameters exploration could result in even better performance than the ones reported in Table 3.

- Hidden size of the classifier : 10 and 50.
- Hidden size of the covariates mapper : 25 and 50.
- Dropout : 0. and 0.1
- Hidden size of the temporal hidden process : 50 and 100.
- Learning rate : 0.001 and 0.005
- Weight decay : 0.0001 and 0.001

GRU-ODE

- Hidden size of the classifier : 20 and 50.
- Hidden size of the covariates mapper : 25 and 50.
- Dropout : 0. and 0.1
- Weight decay : 0.0001, 0.00001 and 0.000001

Appendix F. Pairwise significance of the results

As the objective of this paper is to showcase the significance of using longitudinal information for prognosis in MS, we report in Table F.6 the results of pair-wise tests for the significance of improvement of performance of all pairs of methods. The dynamic RF model is significantly better than the static one. Importantly, all three models using longitudinal measurements are significantly better than the static and dynamic baselines.

p-values	Static	Dynamic	BPTF-SI-RF	GRU-TA	GRU-ODE
Static	/	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**	$<$ **0.001**
Dynamic	$>$ 0.999	/	0.016	0.183	0.469
BPTF-SI-RF	$>$ 0.999	0.984	/	0.104	0.983
GRU-TA	$>$ 0.999	0.817	0.096	0.983	0.249
GRU-ODE	$>$ 0.999	0.531	0.017	0.751	/

Table F.6: Significance of the AUC-ROC results of the different methods. P-values of unilateral pair-wise t-tests are reported. Rows index the reference method to compare against. Columns index the compared method. The reported p-values are computed from testing the alternative hypothesis that the compared method has better performance than the reference method. P-values below a significance threshold of 0.001 are in bold.

Appendix G. Performance per MSCourse

In Table G.7 and on Figure 5 we report the performance of the models split by type of MS. Table G.7 reports the test AUC-ROC and AUC-PR of the different models along with the standard deviations over the 5 folds. Figure 5 shows graphically the AUC-ROC of the different methods (colors) for the different types of MS.

Appendix H. Workflow of the patient cohort definition

In Table H.8, we give the main steps of the cohort definition process that lead to a loss of patients with respect to the original available cohort. The right columns gives the remaining number of patients in the cohort after each step is performed. From $55,409$ patients initially, the cohort contains $6,682$ patients that satisfy the eligibility criteria for our study.

Model	MSCourse	ROC-AUC	AUC-PR
RF-Static	Overall	0.63 ± 0.02	0.26 ± 0.01
	Relapsing-Remitting	0.59 ± 0.02	0.22 ± 0.01
	Primary Progressive (general)	0.47 ± 0.05	0.16 ± 0.05
	Primary Progressive	0.47 ± 0.09	0.16 ± 0.03
	Secondary Progressive	0.55 ± 0.05	0.19 ± 0.05
	Primary Relapsing	0.52 ± 0.25	0.17 ± 0.06
RF-Dynamic	Overall	0.67 ± 0.01	0.25 ± 0.01
	Relapsing-Remitting	0.65 ± 0.02	0.23 ± 0.01
	Primary Progressive (general)	0.52 ± 0.09	0.16 ± 0.03
	Primary Progressive	0.52 ± 0.09	0.16 ± 0.04
	Secondary Progressive	0.56 ± 0.07	0.19 ± 0.06
	Primary Relapsing	0.55 ± 0.14	0.17 ± 0.02
BPTF-SI-RF	Overall	0.68 ± 0.01	0.29 ± 0.01
	Relapsing-Remitting	0.66 ± 0.01	0.27 ± 0.01
	Primary Progressive (general)	0.55 ± 0.09	0.21 ± 0.05
	Primary Progressive	0.54 ± 0.09	0.20 ± 0.05
	Secondary Progressive	0.54 ± 0.04	0.22 ± 0.07
	Primary Relapsing	0.57 ± 0.16	0.18 ± 0.07
GRU-ODE	Overall	0.66 ± 0.02	0.29 ± 0.02
	Relapsing-Remitting	0.64 ± 0.02	0.26 ± 0.02
	Primary Progressive (general)	0.56 ± 0.13	0.24 ± 0.09
	Primary Progressive	0.57 ± 0.15	0.26 ± 0.10
	Secondary Progressive	0.60 ± 0.05	0.29 ± 0.10
	Primary Relapsing	0.59 ± 0.19	0.19 ± 0.07
Time-aware GRU	Overall	0.67 ± 0.01	0.29 ± 0.01
	Relapsing-Remitting	0.65 ± 0.02	0.26 ± 0.02
	Primary Progressive (general)	0.48 ± 0.09	0.20 ± 0.05
	Primary Progressive	0.45 ± 0.09	0.16 ± 0.04
	Secondary Progressive	0.68 ± 0.09	0.27 ± 0.07
	Primary Relapsing	0.58 ± 0.13	0.18 ± 0.11

Table G.7: Results for disability progression prediction per MS Course. Primary Progressive (general) groups primary progressive and primary relapsing together.

Cleaning Step	**Number of patients**
Initial number of patients	55,409
Remove patients with no onset date	55,405
Remove patients with diagnosis date after extraction date	55,381
Remove patients with visit dates after extraction date	55,351
Remove patients with no visits	55,632
Remove patients with no EDSS measurements	51,327
Remove patients with less than 3 visits	35,424
Remove patients with less than 3 years follow-up	23,447
Remove patients with less than 6 visits in the observation window	11,280
Remove patients with less than 1 visit in the test window	8,124
Remove patients with no confirmation EDSS	7,116
Remove CIS patients	6,682
Final Number of patients in the cohort	**6,682**

Table H.8: Description of the patients cohort definition along with the remaining number of patients in the cohort at each step.