THERAPEUTIC ADVANCES in
Neurological Disorders

Original Research

Retrospective cohort study to devise a treatment decision score predicting adverse 24-month radiological activity in early multiple sclerosis

Ther Adv Neurol Disord

2023, Vol. 16: 1–25

DOI: 10.1177/17562864231161892

Alexander Hapfelmeier, Begum Irmak On, Mark Mühlau, Jan S. Kirschke, Achim Berthele, Christiane Gasperi, Ulrich Mansmann, Alexander Wuschek, Matthias Bussas, Martin Boeker, Antonios Bayas, Makbule Senel, Joachim Havla, Markus C. Kowarik, Klaus Kuhn, Ingrid Gatz, Helmut Spengler, Benedikt Wiestler, Lioba Grundl, Dominik Sepp and Bernhard Hemmer

Abstract

Background: Multiple sclerosis (MS) is a chronic neuroinflammatory disease affecting about 2.8 million people worldwide. Disease course after the most common diagnoses of relapsing-remitting multiple sclerosis (RRMS) and clinically isolated syndrome (CIS) is highly variable and cannot be reliably predicted. This impairs early personalized treatment decisions.

Objectives: The main objective of this study was to algorithmically support clinical decision-making regarding the options of early platform medication or no immediate treatment of patients with early RRMS and CIS.

Design: Retrospective monocentric cohort study within the Data Integration for Future Medicine (DIFUTURE) Consortium.

Methods: Multiple data sources of routine clinical, imaging and laboratory data derived from a large and deeply characterized cohort of patients with MS were integrated to conduct a retrospective study to create and internally validate a treatment decision score [Multiple Sclerosis Treatment Decision Score (MS-TDS)] through model-based random forests (RFs). The MS-TDS predicts the probability of no new or enlarging lesions in cerebral magnetic resonance images (cMRIs) between 6 and 24 months after the first cMRI.

Results: Data from 65 predictors collected for 475 patients between 2008 and 2017 were included. No medication and platform medication were administered to 277 (58.3%) and 198 (41.7%) patients. The MS-TDS predicted individual outcomes with a cross-validated area under the receiver operating characteristics curve (AUROC) of 0.624. The respective RF prediction model provides patient-specific MS-TDS and probabilities of treatment success. The latter may increase by 5–20% for half of the patients if the treatment considered superior by the MS-TDS is used.

Conclusion: Routine clinical data from multiple sources can be successfully integrated to build prediction models to support treatment decision-making. In this study, the resulting MS-TDS estimates individualized treatment success probabilities that can identify patients who benefit from early platform medication. External validation of the MS-TDS is required, and a prospective study is currently being conducted. In addition, the clinical relevance of the MS-TDS needs to be established.

Keywords: machine learning, multiple sclerosis, personalized medicine, predictive factor, predictive model, treatment effect

Received: 19 August 2022; revised manuscript accepted: 19 February 2023.

Correspondence to:
Alexander Hapfelmeier
Institute of AI and Informatics in Medicine, School of Medicine, Technical University of Munich, Ismaninger Str. 22, Munich 81675, Germany.

Institute of General Practice and Health Services Research, School of Medicine, Technical University of Munich, Orleansstr. 47, Munich 81667, Germany.

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany
alexander.hapfelmeier@mri.tum.de

Begum Irmak On
Ulrich Mansmann
Institute for Medical Information Processing, Biometry, and Epidemiology, Ludwig-Maximilians-Universität in Munich, Munich, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Mark Mühlau
Achim Berthele
Christiane Gasperi
Alexander Wuschek
Matthias Bussas
Department of Neurology, Klinikum rechts der Isar School of Medicine, Technical University of Munich, Munich, Germany

Jan S. Kirschke
Benedikt Wiestler
Lioba Grundl
Dominik Sepp
Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

Martin Boeker
Klaus Kuhn
Ingrid Gatz
Helmut Spengler
Institute of AI and Informatics in Medicine, School of Medicine, Technical University of Munich, Munich, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Antonios Bayas
Department of Neurology, Medical Faculty, University of Augsburg, Augsburg, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Makbule Senel
Department of Neurology, Ulm University Hospital, Ulm, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Joachim Havla
Institute of Clinical Neuroimmunology, LMU Hospital, Ludwig-Maximilians-Universität in Munich, Munich, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Markus C. Kowarik
Department of Neurology & Stroke and Hertie-Institute for Clinical Brain Research, Eberhard-Karls University of Tübingen, Tübingen, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Bernhard Hemmer
Department of Neurology, Klinikum rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany

Data Integration for Future Medicine (DIFUTURE) Consortium, Munich, Germany

Munich Cluster for Systems Neurology (SyNergy), Munich, Germany

Introduction

Multiple sclerosis (MS) is a chronic neuroinflammatory disease affecting more than 200,000 people in Germany and 2.8 million people worldwide.[1,2] At the time the disease becomes symptomatic, it is classified as clinically isolated syndrome (CIS), relapsing-remitting multiple sclerosis (RRMS), or primary progressive multiple sclerosis (PPMS). CIS is a patient's first clinical event without meeting criteria of dissemination both in time and space.[3] Many patients with CIS are likely to convert to RRMS later on. A number of disease-modifying therapy (DMT) options have been approved and are available for treatment of patients with RRMS and CIS. Treatment with DMT is most efficacious during the early phase of the diseases and the efficacy of DMTs decreases over time especially when the disease converts into secondary progressive MS.

Although many patients take advantage of early DMT, long-term studies have demonstrated that a proportion of patients with CIS and MS, who are not treated with DMT, do not acquire significant disability even decades after diagnosis.[4] Given the increase in prevalence over the last decades[5] and the observation of a much better prognosis of recently diagnosed patients, which cannot be fully explained by the availability of DMT,[6] it is conceivable to conclude that a subset of patients with MS or CIS may not require long-term DMT treatment. Identifying these patients may not only protect them from possible side effects of DMTs, which often go along with impaired quality of life, but may also avoid significant costs for the health care system. Thus, the algorithms that allow to stratify patients with respect to prognosis and treatment responses are warranted.

The course of the disease is difficult to predict from the onset and varies greatly among patients. Therefore, various data sources have been used to identify prognostic factors and build multivariable predictive models for disease progression through statistical modelling and machine learning. The results of several recent systematic reviews show that there is a broad awareness of the relevance of the research question, which is reflected in the extensive literature and the many proposed prognostic models.[7–9] Their common conclusion, however, is that most of the reviewed studies and respective models are at high risk of bias and lack external validation. A few methodologically well-conduced examples with low risk of bias exist, but these models show only weak accuracy.[10–12] A related and even more complex research question arises from the field of personalized medicine and concerns the modification of treatment effects by predictive factors. Appropriate predictive models should support individualized treatment recommendations based on a patient's characteristics. As a practical consequence, patients requiring effective treatment at an early stage could be identified, as well as patients with an expected mild disease course who may not be unnecessarily exposed to the risk of adverse effects.

This retrospective monocentric cohort study (Retro-MS) was conducted to develop and internally validate a clinically relevant and individualized treatment decision score (Multiple Sclerosis Treatment Decision Score – MS-TDS) for newly diagnosed CIS and RRMS patients. The MS-TDS is supposed to support the treating physician and the patient in making an informed decision based on anticipated treatment success between no or platform medication. A further objective was to identify patient features from clinical, imaging and laboratory data that are predictive factors in this regard.

Methods

The Retro-MS cohort was formed from the routine care patients treated at the Department of Neurology at the Klinikum rechts der Isar of the Technical University of Munich (TUM) to create the MS-TDS by predictive modelling of existing multidimensional baseline data. The MS-TDS predicts the outcome of no new or enlarging T2-lesions[13–16] in cerebral magnetic resonance images (cMRIs) of a newly diagnosed CIS or RRMS patient on platform or no medication between 6 and 24 months after their first cMRI using patient features collected at baseline. This outcome is considered to be a sensitive short-term surrogate of clinical disease activity observed in the long-term.[17] Baseline was defined as the first date of cMRI available or DMT start date, whichever occurred first. Patients were followed-up as long as they had eligible cMRI, which is defined as a cMRI acquired until 32 months after baseline or until the first cMRI acquired after 32 months.

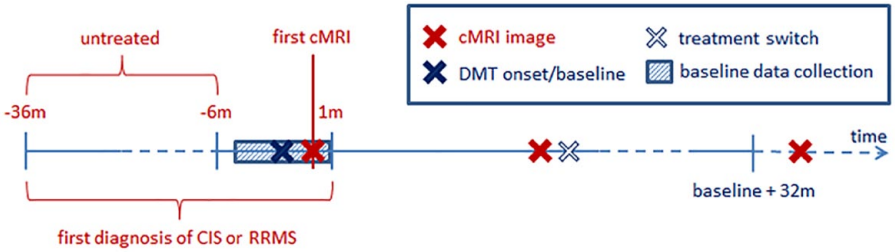

Figure 1. Patient-level data progression for an example patient.

Study population and sample

Patients treated at the TUM Department of Neurology at the Klinikum rechts der Isar during the years 2008–2017 were taken into account. They were diagnosed according to the 2005 and 2010 McDonald diagnostic criteria depending on the time point of diagnosis. They were all seen in the outpatient centre of the department in regular intervals, and clinical parameters related to disease activity and severity were recorded. To align the sample of Retro-MS with the study population of the ongoing prospective validation study ProVal-MS,[18] we applied the following selection criteria. Only patients diagnosed with CIS or RRMS earliest 3 years before and latest 1 month after their first cMRI were included. They also had to be previously untreated, including the possibility of a DMT no earlier than 6 months before their first cMRI. Thereby, a period of 6 months was considered as the run-in phase after which a medication becomes effective.[19] Patients with less than two cMRI images available or a difference of more than 32 months between their consecutive cMRI images were excluded. Patients whose data complied with the above rules were eligible for analysis. An illustration of these definitions is given in Figure 1 for an example patient.

Platform medication included the following DMTs: Glatiramer acetate, Interferon-beta 1a and 1b, Peg-Interferon, Dimethyl fumarate and Teriflunomide. A total of 12 patients, who received a more active DMT (e.g. Alemtuzumab, Cladribine, Natalizumab, Mitoxantrone, Rituximab) as first medication, were excluded from analysis. Each interval between two consecutive cMRI dates was assigned one of the two treatment regimens of no medication or platform medication, depending on which was used for the majority of the time within that interval.

Data

Because Retro-MS is based on routine clinical data, the timings of clinical assessments were not under the control of the investigators. Many feature values were collected at baseline, that is, within a 6-month time window around the first date of cMRI or DMT onset. If any data or measurements existed during the 3-month period prior to this date, then the latest of those measurements was considered the baseline value. Otherwise, the earliest measurement within the 3-month period after this date was considered the baseline value (Figure 1). Standarized cMRIs were available from the beginning of 2009 until the end of 2017. Baseline data were collected from 2008 onwards and the outcome assessment was limited to until the end of 2017. The timeframes and definitions of the features included in the analysis were consented between the centres participating in the Retro-MS and ProVal-MS studies. Details are provided in Appendix 1.

Data were exported from different clinical information systems as specified below to a central staging area in the protected clinical network. Data were extracted from tabular form in the staging area and loaded into the Informatics for Integrating Biology and the Bedside (i2b2) and TranSMART data marts for data exploration. Final data integration, data cleaning and construction of patient histories were performed within the software R version 3.6.3 (The R Foundation for Statistical Computing, Vienna, Austria) by creating a data frame object that was used for analysis. All steps were performed according to German and European data protection regulations.

Clinical data. Clinical data were collected during outpatient visits of the patients and stored in the clinical information system. These included

demographics, information on diagnosis and clinical presentation at onset, occurrence and clinical presentation of relapses, disease severity [Expanded Disability Severity Scale (EDSS), multiple sclerosis functional composite (MSFC)], fatigue [Fatigue Scale for Motor and Cognitive Functions (FSMC)] and depression [Beck Depression Inventory (BDI)].

Imaging data. The cMRIs were acquired during routine clinical practice at one and the same 3Tesla scanner (Achieva; Philips Healthcare, Best, the Netherlands) and stored in the radiology information system. The intervals between available consecutive images were of different length. The respective outcome assessment was performed retrospectively in a semi-automated manner based on a fluid-attenuated inversion recovery (FLAIR) sequence [voxel size = $1.5 \times 1 \times 1$ mm; repetition time (TR) = 10,000 ms; time to echo (TE) = 140 ms; inversion time (TI) = 2750 ms] and a three-dimensional (3D) spoiled gradient echo T1-weighted sequence (1 mm isotropic; TR = 9 ms; TE = 4 ms). All images were converted from dicom to Nifti format using dcm2niix. First, lesions in baseline scans were automatically segmented by the lesion segmentation tool (LST, https://www.applied-statistics.de/lst.html) yielding binary lesion segmentations in native space.[20] Next, all images were rigidly co-registered to the T1-weighted image of the same time point using NiftyReg. Then, all images were rigidly brought to Montreal Neurological Institute (MNI) space, and skull-stripped using parameters derived from the T1-weighted image of the same time point (HD-BET, github.com/MIC-DKFZ/HD-BET).[21] Now, segmented lesions from baseline images were labelled according to their location (periventricular, juxtacortical/cortical, infratentorial, subcortical/unspecific) using an atlas-based approach, in which the MNI tissue atlas was deformably registered onto the T1-weighted image using ANTs SyN. Segmented lesions were manually reviewed and corrected by one out of four experienced neuroradiologists using ITK-SNAP.[22] Baseline FLAIR images were rigidly co-registered to follow-up FLAIR images using NiftyRegand, to ensure comparable image intensities, FLAIR baseline images were intensity-scaled according to FLAIR follow-up images by a histogram-matching algorithm (using the 'match_histograms' function of the Python package scikit-image); finally, subtraction images were rendered by a voxel-wise subtraction of the baseline FLAIR image from the follow-up FLAIR image. In these difference images, raters only segmented new or enlarging lesions.[23] New solitary lesions had to be at least 3 mm in diameter according to the current diagnostic criteria.[3] New lesions that showed any overlap (i.e. >0 voxels) with an existing lesion and that, hence, could be regarded enlarged were counted if the new lesion area was of a shape that (virtually) could best be described by two (or even more) spheroids, as we then assumed that a new lesion had grown into an existing one. Again, only those lesions with an estimated diameter of at least 3 mm were counted. Lesions having enlarged along the whole of their circumference (towards brain parenchyma) were only counted if the enlargement was clear to the observers. Such 'truly' enlarged lesions were hardly ever observed. Both new and enlarging lesions were considered as disease progression. All image evaluations were finally reviewed by one senior neuroradiologist (J.S.K.). This assessment of lesions was blinded to the treatment, medication or future cMRI of a patient.

Laboratory data. Routine laboratory data were generated by the central clinical laboratory and stored in the laboratory information system. Cerebrospinal fluid (CSF) data were generated by the CSF laboratory of the Department of Neurology. Data were transferred into the clinical information system.

Statistical analysis

In Retro-MS, the definition of baseline data and outcome assessment is mainly governed by the timing of cMRI. The latter, however, is not perfectly regular due to the fact that the patient visits in routine care can deviate from preplanned schedule and observations may have different patterns for different patients. This obstacle was overcome by conceptualizing predictive modelling in a time-to-event framework. The occurrence of the primary endpoint – that is, new or enlarged cMRI lesions between consecutive images – was considered as an interval-censored event. The conditional probability of observing an event time T between 6 and 24 months for a patient with feature vector X, given the event did not occur until month 6, calculates to $P(6 < T < 24 \mid T > 6, X) = P(6 < T < 24 \mid X) / P(T > 6 \mid X) = (S(6 \mid X) - S(24 \mid X)) / S(6 \mid X)$. The complement of this conditional probability, that is, $1 - P(6 < T < 24 \mid T > 6, X) = P(T > 24 \mid T > 6, X)$,

is the MS-TDS that predicts treatment success, instead of treatment failure. In the formula above, $S(t \mid X)$ is the probability of no event until time t for a patient with feature vector X. In this framework, we assumed noninformative censoring.

A predictive RF model was implemented through transformation forests based on fully parameterized Cox proportional hazards models (using a smooth baseline hazard function) to deal with the interval-censored outcome and to finally provide the MS-TDS.[24,25] The predictive RF had treatment (no medication *versus* platform medication) as a predictor variable in the underlying Cox models while other features were used as potential splitting variables to build the tree structure of the forest. With this approach, the interaction of the features with treatment is explicitly modelled. The optimized hyperparameters of the RF were the number of variables randomly sampled as candidates for splitting (usually termed 'mtry') and the minimum number of observations to be considered for splitting ('minsplit'). The hyperparameter tuning was based on a prespecified set of potential values. Recommended values $\sqrt{p}$ and $p/3$ were chosen for mtry, in which p is the number of splitting variables.[26] The set of values considered for minsplit was 20, 30 and 40. The combination of these values resulted in six models for the model selection procedure described below. The implementation of the RF was through the R packages 'trtf' and 'tram'.[24,27,28] Package defaults were used for the remaining hyperparameters.

A benchmark study was performed for hyperparameter tuning and to choose the best performing model as well as to obtain an unbiased estimate of its performance. The area under the receiver operating characteristics curve (AUROC) at 24 months served as the corresponding performance measure using the MS-TDS as predictor variable.[29] Models were compared by internal validation, that is, *via* nested threefold cross-validation. Thereby, a best performing model was determined in each inner cross-validation loop. These models were refit to the whole data of the respective inner loop and applied to the test data of the corresponding outer loop to obtain unbiased performance estimates. The average of these values provides an unbiased assessment of the overall performance of a best model. The best model itself was determined through the best performing model in the outer loop and refit to the whole data to produce the MS-TDS. Likelihood-based permutation variable importance measures (VIMPs) of this final model were used to identify informative predictor variables.[30,31] Predictor variables with a VIMP lower than the VIMP of an additionally included random variable were excluded from VIMP display.[32] To further evaluate the counterfactual analysis, the MS-TDS was calculated assuming both treatment alternatives for each patient and compared between the actual medication groups.

In addition to the above-mentioned analyses, baseline data were described for the analysis cohort by medication group. Descriptive statistics used are absolute and relative frequencies for categorical variables and median and interquartile range (IQR) for numeric or ordinal variables. The interval-censored outcome was described by plotting Weibull estimates of event probabilities by medication group.

Before the analysis, missing values of patient features were imputed by an RF imputation model provided by the R package 'missForest'.[33] The interval-censored outcome was omitted during imputation to prevent artificially creating relations between that and the patient features.

All analyses were performed with the software R 3.6.3 (The R Foundation for Statistical Computing). The session info including information about the used packages is provided in Appendix 2. A Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) checklist for prediction model development and validation is provided in Appendix 4.[34]

Results

A total of 2992 patients had a record of any neurological diagnosis in the clinical data of the neurology department during the years 2008–2017. Of these, 774 patients had a diagnosis of CIS or RRMS and at least two T2 FLAIR cMRI images. A subset of 509 patients further met the eligibility criteria based on diagnosis, imaging and treatment history. A further 34 patients had to be excluded because of technical problems in image analysis, pregnancy, unknown medication status or use of high efficacy medication from baseline. Finally, 475 patients contributed 1804 images to

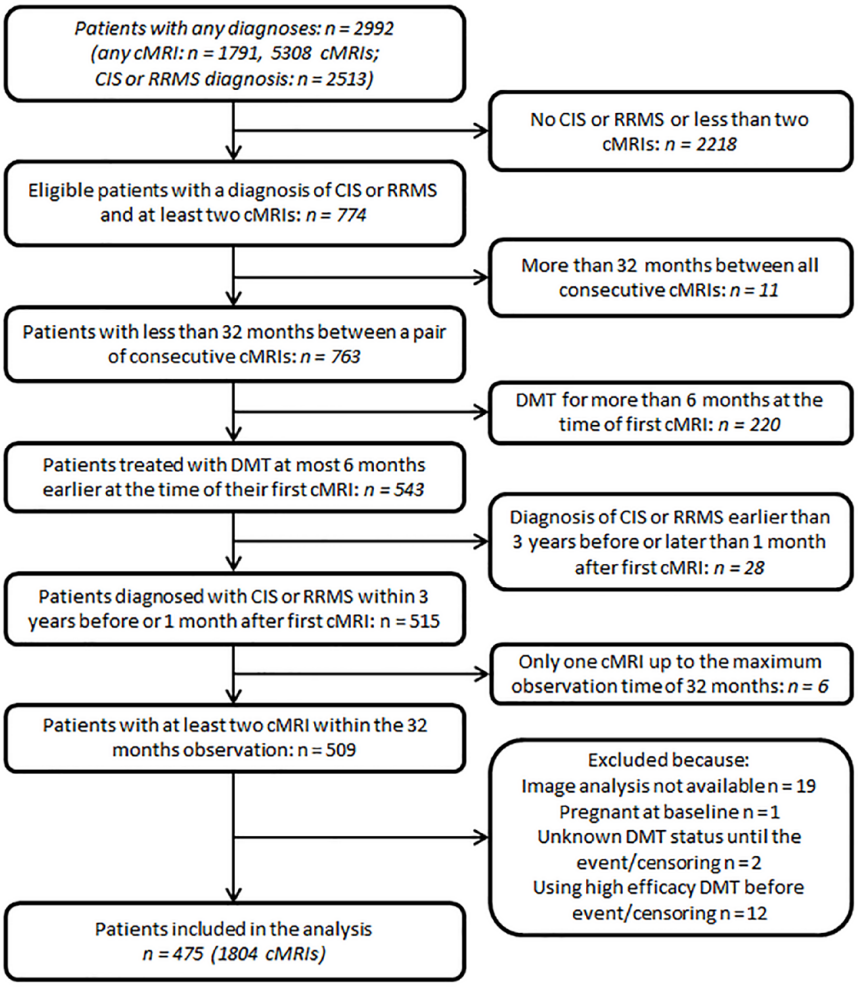

Figure 2. Flow chart of patient selection.

analysis. A detailed flow chart of patient selection is presented in Figure 2.

A summary of baseline characteristics by medication is given in Table 1. No medication and platform medication were administered to 277 (58.3%) and 198 (41.7%) patients at baseline, respectively. Patients with no medication at baseline had fewer lesions in their first cMRI (median 13.0 *versus* 17.5), and were more likely to be diagnosed with CIS rather than RRMS (54.2% *versus* 45.5%). The primary endpoint was met by 214 patients and 167 patients under no medication and platform medication, respectively. Some data were missing in 44 of 65 features (68%) with an average of 20.5% missing values per feature (median = 22.5%, IQR = 0.0–36.0%). A detailed presentation of the number of missing values per feature and medication group is given in Table 1.

The estimated probabilities of observing no event until time t are displayed in Figure 3. The estimated median time-to-event is 122.0 days (4.0 months) under platform medication and 136.9 days (4.5 months) under no medication. The estimated probability of being event-free at 6 months is 43.2% and 45.3% under platform medication and no medication, respectively. At 24 months, these probabilities are 18.6% and 20.5%, respectively.

The average performance of the best prediction model was AUROC = 0.624 (in-depth information about the results of the benchmark study are given in Appendix 3). The VIMPs of the most important features of the final model, that exceeded the VIMP of a random noise variable, are displayed in Figure 4. Medication is clearly the most important predictor variable. Demographics

Table 1. Baseline characteristics of all patients by medication at baseline.

	No medication (*n* = 277)			Platform medication (*n* = 198)		
	Missing values	**Median/*N***	**IQR/%**	**Missing values**	**Median/*N***	**IQR/%**
Outcome-related						
Follow-up since first cMRI (days)	0	365	(175–623)	0	378.5	(359.0–680.5)
Primary endpoint	0	214	(77.3%)	0	167	(84.3%)
Demographics						
Sex (F)	0	184	(66.4%)	0	134	(67.7%)
Age (years)	0	34.7	(28.1–42.0)	0	32.4	(25.6–39.3)
Height (cm)	110	170	(164–178)	33	172	(165–178)
Weight (kg)	110	72.5	(62.5–85)	33	73	(62.5–85)
BMI (kg/m^2)	110	24.2	(22.0–27.7)	33	24.2	(21.1–27.9)
Smoking (yes/no/former/unknown)	0	86/98/45/48	31.0/35.4/16.2/17.3%	0	71/82/34/11	35.9/41.4/17.2/5.6%
Diagnosis						
Diagnosis at baseline (CIS)	0	150	(54.2%)	0	90	(45.5%)
Time since diagnosis at baseline (days)	0	31	(16–76)	0	54	(25.2–78.8)
First symptom						
Numbness	0	142	(51.3%)	0	104	(52.5%)
Other cranial nerve symptom	0	47	(17.0%)	0	22	(11.1%)
Paresis	0	32	(11.6%)	0	20	(10.1%)
Optic neuritis	0	91	(32.9%)	0	62	(31.3%)
Any other symptom	0	54	(19.5%)	0	51	(25.8%)
Relapses (± 3 months from baseline)						
Numbness	0	113	(40.8%)	0	82	(41.4%)
Other neurological symptom	0	45	(16.2%)	0	22	(11.1%)
Paresis	0	29	(10.5%)	0	19	(9.6%)
Optic neuritis	0	65	(23.5%)	0	48	(24.2%)
Any other symptom	0	56	(20.2%)	0	56	(28.3%)

(Continued)

Table 1. (Continued)

	No medication (*n* = 277)			Platform medication (*n* = 198)		
	Missing values	**Median/*N***	**IQR/%**	**Missing values**	**Median/*N***	**IQR/%**
Number of relapses during 3 years before baseline	0	1	(1–1)	0	1	(1–2)
EDSS/functional score						
Total	64	1	(0–2)	35	1	(1–2)
Pyramidal	145	0	(0–1)	94	0	(0–1)
Cerebellar	153	0	(0–0)	102	0	(0–1)
Brainstem	150	0	(0–0)	102	0	(0–0)
Sensory	137	0	(0–1)	98	1	(0–1)
Bowel and bladder	159	0	(0–0)	104	0	(0–0)
Visual	135	0	(0–1)	92	1	(0–1)
Cognitive	157	0	(0–0)	105	0	(0–0)
Ambulation	223	0	(0–0)	167	0	(0–0)
FSMC						
Fatigue cognitive	124	15	(11–22)	56	14.5	(10–24)
Fatigue motor	124	16	(12–23)	56	16	(11.2–25.0)
Total	125	32	(23.0–47.2)	60	31	(23.0–48.8)
MSFC						
Nine-Hole Peg test result – hand/arm	88	17.6	(16.1–19.0)	36	17.5	(16.1–19.0)
25-Foot Walk test result – ambulation	88	4	(3.4–4.5)	37	3.9	(3.5–4.4)
BDI-II – depression	88	5	(1–9)	37	6	(2–9)
First cMRI						
Total lesions count	0	13	(5–27)	0	17.5	(9–37)
Periventricular lesions present	0	243	(87.7%)	0	187	(94.4%)
Subcortical/unspecific lesions present	0	249	(89.9%)	0	189	(95.5%)
Juxtacortical or cortical lesions present	0	188	(67.9%)	0	157	(79.3%)

(Continued)

Table 1. (Continued)

	No medication (*n* = 277)			Platform medication (*n* = 198)		
	Missing values	**Median/*N***	**IQR/%**	**Missing values**	**Median/*N***	**IQR/%**
Infratentorial lesions present	0	125	(45.1%)	0	103	(52.0%)
CSF						
Leucocyte count (cnt/mcl)	100	6	(3–12)	76	7	(3.2–12)
Glucose (mg/dl)	100	62	(57–68)	76	63	(57–71)
Total protein (mg/dl)	98	491	(398–637)	75	532	(428–635)
Albumin quotient ($\times 10^{-3}$)	96	5.3	(3.9–7.0)	75	5.4	(4.3–7.0)
CSF-specific oligoclonal bands (no/borderline/yes)	95	21/29/132	11.5/15.9/72.5%	75	8/21/94	6.5/17.1/76.4%
IgG Quo/Alb Quo (IgG-index)	96	0.7	(0.5–1.0)	75	0.7	(0.5–1.0)
IgM Quo/Alb Quo (IgM-index)	107	0.1	(0.1–0.1)	79	0.1	(0.0–0.2)
IgA Quo/Alb Quo (IgA-index)	96	0.3	(0.2–0.3)	76	0.3	(0.2–0.3)
Laboratory – blood analysis						
Basophils (%)	81	0	(0–1)	50	0	(0–1)
Bilirubin (mg/dl)	81	0.5	(0.4–0.7)	51	0.5	(0.4–0.7)
Blood urea nitrogen (mg/dl)	60	13	(10–15)	37	13	(11–15)
Eosinophils (%)	76	1	(1–2)	44	1	(1–2)
Erythrocytes (cnt/pl)	30	4.6	(4.4–5.0)	9	4.6	(4.3–4.9)
GOT (ASAT) (U/l)	94	25	(21–31.5)	29	25	(21–31)
GPT (ALAT) (U/l)	32	22	(16–34)	9	23	(17–36)
Haematocrit (%)	30	40.6	(38.3–43)	9	40.4	(38.3–43.5)
Haemoglobin (g/dl)	30	13.9	(13.0–14.9)	9	13.9	(13.1–15.0)
Leucocytes (cnt/nl)	30	7	(5.7–9.0)	9	6.9	(5.7–9.8)
Lymphocytes (%)	73	28	(20–35)	42	26	(21–33)
MCH (pg)	30	30	(29–31)	9	30	(29–31)

(Continued)

Table 1. (Continued)

	No medication ($n = 277$)			Platform medication ($n = 198$)		
	Missing values	**Median/*N***	**IQR/%**	**Missing values**	**Median/*N***	**IQR/%**
MCHC (g/dl)	30	34.4	(33.7–35.1)	9	34.3	(33.6–34.9)
MCV (fl)	30	88	(85–90)	9	88	(86–91)
Monocytes (%)	73	7	(6–9)	42	7	(6–9)
Neutrophils (%)	87	63	(54–70)	54	64.5	(58–70)
Thrombocytes (cnt/nl)	30	245	(215.5–285.5)	9	242	(204–277)
TSH (mcIU/ml)	79	1.7	(1.2–2.3)	45	1.7	(1.2–2.2)

AST, aspartate transaminase; ALAT, alanine transaminase; BDI, Beck Depression Inventory; BMI, body mass index; CIS, clinically isolated syndrome; cMRI, cerebral magnetic resonance images; CSF, Cerebrospinal fluid; EDSS, Expanded Disability Severity Scale; FSMC, Fatigue Scale for Motor and Cognitive Functions; GOT, glutamic oxaloacetic transaminase; GPT, glutamic-pyruvic transaminase; IgA, Immunoglobulin A; IgG, Immunoglobulin G; IgM, Immunoglobulin M; IQR, interquartile range; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular haemoglobin concentration; MCV, mean corpuscular volume; MSFC, MS functional composite; Quo, quotient; TSH, thyroid-stimulating hormone.

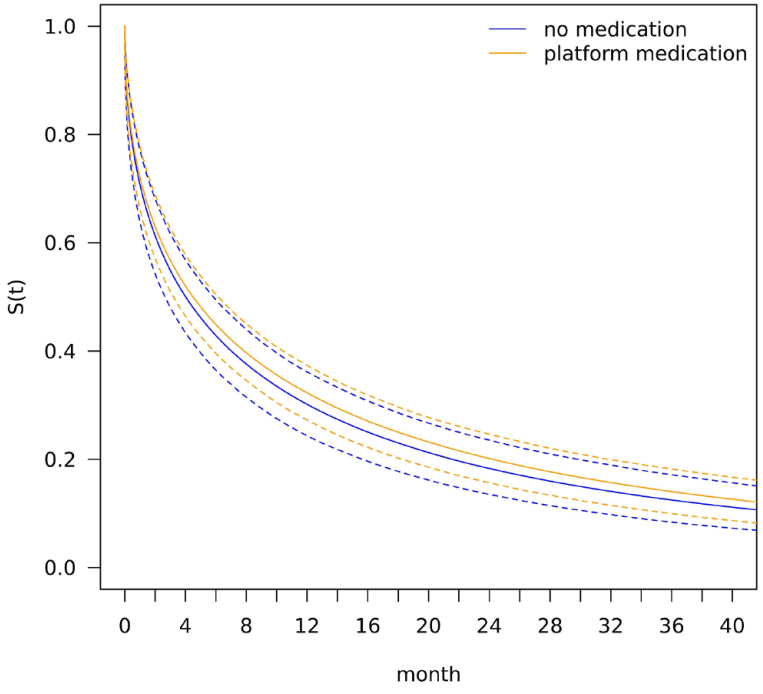

Figure 3. Weibull estimates (solid lines) and pointwise 95% confidence intervals (dashed lines) of observing no event until time *t*, that is, estimates of *S*(*t*), by medication at baseline.

like age, weight, height, body mass index (BMI) and sex also play an important role in the prediction of the outcome and as predictive factors. Further relevant patient features include the lesion count at baseline as well as the presence of periventricular lesions, the number of relapses, the diagnosis at baseline (CIS or RRMS) as well as the presence of CSF-specific oligoclonal bands and interestingly the Immunoglobulin A (IgA)-index (i.e. IgA Quotient/Albumin Quotient). It is important to note that causal effects or biological relevance cannot be inferred for the variables listed in Figure 4, as the prediction model can also benefit from spurious correlations with the outcome.

Predicting the outcome

The final model provides the MS-TDS, which is the probability of observing no new or enlarged T2-lesion between 6 and 24 months under platform medication or no medication. Given a patient's characteristics X and event time T, it is defined as $P(T > 24 \mid T > 6, X)$. In addition, the conditional probability of observing no event until time t, given the event did not occur before month 6 can be calculated by $P(T > t \mid T > 6, X) = P(T > t \mid X) / P(T > 6 \mid X) = S(t \mid X) / S(6 \mid X)$ with $t \geq 6$. Some illustrative examples are given in Figure 5. Each figure shows two curves, one for each medication group. This is a counterfactual information for the treating neurologist regarding what would happen under no medication or platform medication in terms of radiologically assessed disease activity. The

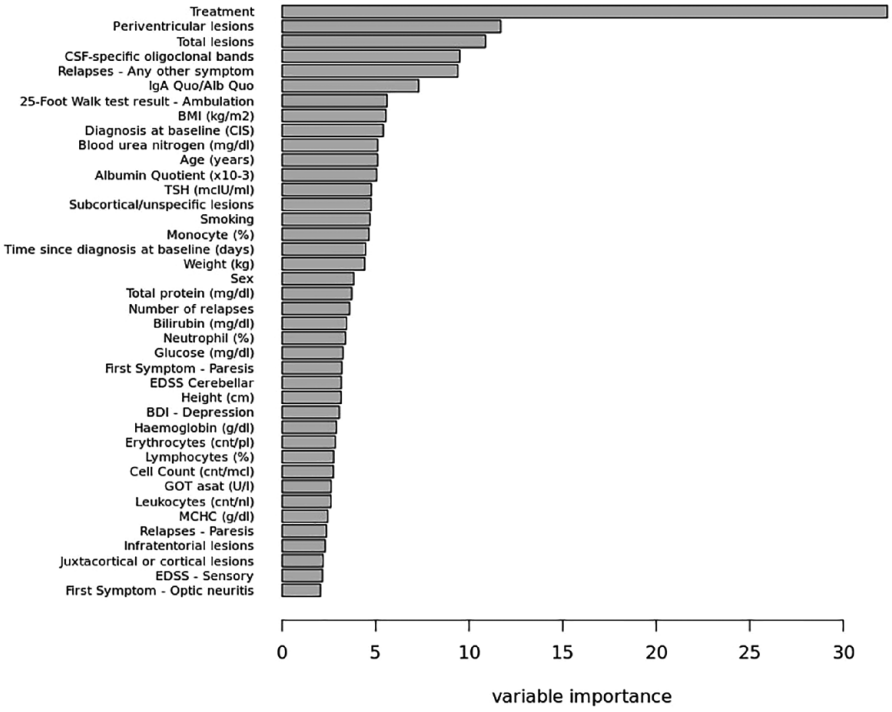

Figure 4. Predictor variables with VIMP exceeding the VIMP of a random noise variable in the final model.

treatment with the higher curve provides the best prognosis in terms of no radiological disease activity for the particular patient. Given the patient features at baseline, the treatment with the better prognosis could be recommended for the patient. For example, a patient may be expected to benefit from platform medication (cf. Figure 5(a) *versus* (b) and (c)).

According to the MS-TDS, about 61.4% of patients with no medication would have benefitted from platform medication with an expected median increase of 6.9% (IQR = 3.7–10.9%) in the probability of being event-free between month 6 and month 24. For patients with platform medication, it is expected that 45.5% benefitted from this treatment option by a corresponding 5.1% (IQR = 2.3–12.8%). These and additional numbers, as well as the distribution of expected differences in the probabilities of being event-free between month 6 and month 24 under either potential treatment option, are shown in Figure 6. In summary, the median and maximum values suggest that for half of the patients, the risk of an event is expected to be reduced by about 5–20% if the treatment recommended by the TDS is given.

Discussion

Data collected in routine practice are becoming increasingly available for observational studies in MS research on predictive factors and related treatment decisions.[35] Similarly, advanced statistical and machine learning methods are continuously evolving within a theoretically sound framework for estimating average and individual treatment effects (ITEs).[25,36–38] Building on these insights, we developed and internally validated the MS-TDS to predict the outcome of no new or enlarging cMRI lesions in a newly diagnosed CIS or RRMS patient on platform or no medication between 6 and 24 months after their first cMRI.

We are publishing the results from the MS-TDS development before performing an external validation study. This provides an objective statement on what will be evaluated in the planned external validation based on the ProVal-MS cohort (ProVal-MS study; German Clinical Trails Register study ID: DRKS00014034).

A predictive RF based on fully parameterized Cox proportional hazards models was fit to the prediction problem and internally validated in a

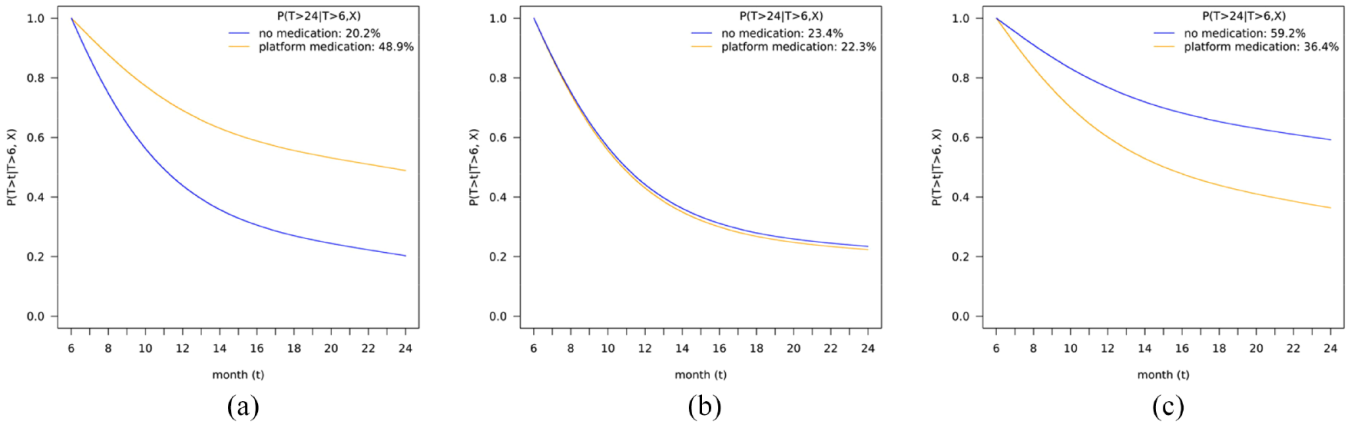

Figure 5. Illustration of the individual probabilities of being event-free, given the event did not occur before month 6, that is, $P(T > t \mid T > 6, X)$, as predicted by the final model. The respective MS-TDS, which corresponds to the probability of observing the event later than after 24 months, that is, $P(T > 24 \mid T > 6, X)$, is given in the figure legends. Each figure shows a patient who benefits from no medication (a), platform medication (c) or neither option (b).

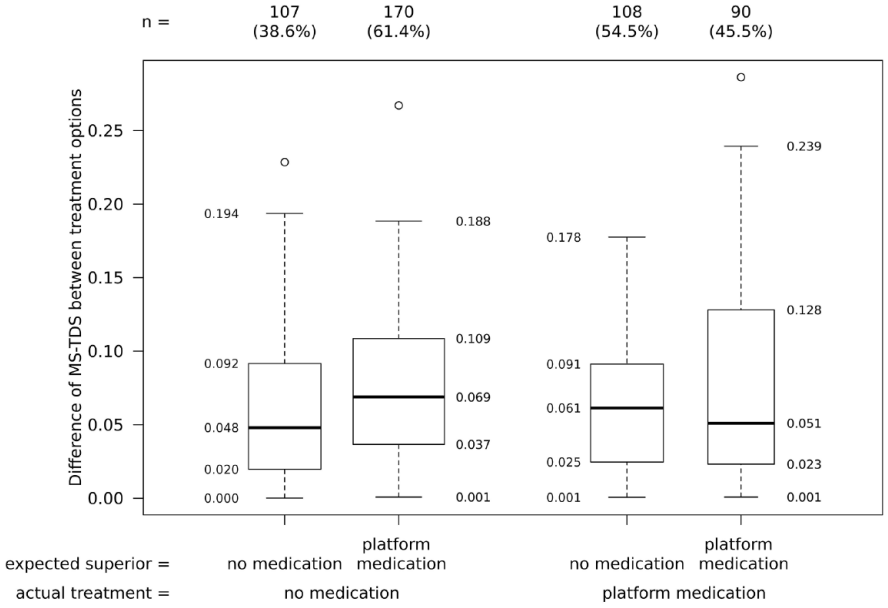

Figure 6. Increase in the probability of being event-free between month 6 and month 24 if the treatment option considered superior according to the MS-TDS rather than the inferior one would have been or was administered to a patient, stratified by actual treatment.

benchmark study that included hyperparameter tuning. The resulting MS-TDS is informative in many ways, suggesting features relevant to the prediction problem by calculating variable importance measures, and predicting individual patient probabilities of being event-free, as a function of time or for the focused time frame of 6–24 months. An illustration of the latter showed that it is possible to identify newly diagnosed patients who would benefit from no medication or platform medication through computation of the MS-TDS based on individual characteristics.

This study identified a number of clinical, laboratory and imaging features as predictive factors of the investigated outcome. Among the strongest

predictor variables were the total lesion count at baseline as well as the presence of periventricular lesions, the diagnosis at baseline (CIS *versus* RRMS), the number of relapses before baseline as well as two CSF parameters – that is the presence of CSF-specific oligoclonal bands and the IgA-index. Most of these have previously been identified as being predictors of the disease course of MS.[39–41] In contrast to previous studies, our model is based on an unbiased approach without preselecting supposedly informative variables and internally validated. To our knowledge, none of the prediction models reported so far had low risk of bias in their model development and evaluation steps and performed higher than area under the curve (AUC) = 0.7.[10,11,12,42] Although the performance of the model is weak, the results of the prediction are promising and provide a basis for future developments.

The present Retro-MS study has several limitations. The analysis of observational data obtained from nonrandomized trials for treatment effect estimation and the exploration of predictive factors generally inherits the risk of confounding and selection bias. The inferiority of platform medication illustrated in Figure 3 indicates that this may also apply to this study. A counterfactual framework involving the concept of potential outcomes, which are the expected outcomes of a patient under each treatment, has been suggested as a solution and has been applied in the present work. A commonly used approach is the application of weighted, conditional or stratified analyses to estimate average treatment effects (ATEs). An underlying assumption is the strongly ignorable treatment assignment (SITA), which suggests that the actual treatment assignment is conditionally independent from the potential outcomes given the observed covariables. Against that background, potential outcomes can be estimated from the observed data.[36] A known limitation of the potential counterfactual framework, which is a limitation shared with any other study trying to estimate ATE or ITEs from observational data of nonrandomized trials, is that SITA might not hold. This might result in biased effect estimation and models. Such models, however, might still be useful for prediction purposes, which is a property that was internally validated in this study. Another source of potential selection bias is the fact that the data were collected at a specialized centre. External validation will be provided by the subsequent prospective and multicentric ProVal study and may indicate such problems.

Lu *et al.*[36] suggest the estimation of ITE by RF under SITA. In a comparison of several RF implementations, they found that tuned RF, with a separate RF model fit to each treatment group, performed best. The strategy of fitting separate models to the treatment groups has been criticized though. Powers *et al.*[37] state, for the case of two treatments, that 'it is to be expected that the selected basis be different between the 2 regression functions. This can cause differences between the conditional means attributable not to a heterogeneous treatment effect but rather to randomness in the basis selection'. In this study, we therefore fitted tuned RF to the whole data and estimated treatment effects within Cox models simultaneously fitted to both treatments.

Furthermore, the reality of routinely collected data necessitated defining the outcome as interval censored, dealing with missing values, and making consensus decisions regarding ambiguous data. Even at the level of feature definitions, strong assumptions had to be made to consider some features as 'none' when there was no entry in the source data. Such decisions were made in consensus meetings with the authors. The assignment of baseline measurements and of treatment groups to the intervals had to be operationalized. In combination with the methodologically challenging task to properly estimate treatment effects, these conditions narrowed the set of applicable statistical models and machine learning methods to a model-based RF, a recently developed one used in the present work. In addition, only internal validation using patients from the same clinic and period could be performed with the available data set. The results on predictive factors should be considered exploratory as they were only discovered to be relevant during the analysis, although the set of potential features was determined in preparatory consensus meetings of the investigators. The routine data also carry a potential risk of misclassification, for example, in the diagnosis of CIS, where oligoclonal band analysis was not recorded in 174/240 (72.5%) patients. The true misclassification rate, however, is likely to be lower because the diagnosis of CIS was based on further diagnostic criteria depending on the time of diagnosis. A similar problem is posed by unobserved confounding, which may be present and may have led to biased findings. Another source of potential bias is the selection of the study population with the imposed restrictions on data availability and timing of DMT

onset, first cMRI and diagnosis. For these reasons, the aforementioned prospective ProVal-MS study was initiated simultaneously to the present Retro-MS study to allow an external and unbiased assessment of the MS-TDS. With an AUROC of 0.624, the performance of the MS-TDS can be considered weak[43] but is comparable to the performance of other models from studies with low risk of bias.[10,11,12,42] The robustness of the result has yet to be demonstrated in a prospective study currently being conducted for external validation.[18] Clinical relevance also remains to be proven. While the score can provide clear treatment recommendations, there are also patients for whom no clear decision is possible (see Figure 5(b)). Some variables that have shown to be relevant in the present prediction model may be so because of spurious correlations, and may not be of direct biological relevance to the outcome. Further insights into more specific medication subgroups with differential efficacy – such as dimethyl fumarate, teriflunomide and injectable medications – could not be obtained due to the moderate overall sample size available for analysis in this study. Further targeted studies with increased sample sizes are needed to investigate such differences and provide more specific treatment recommendations.

Conclusion

Clinical routine data can be used to support treatment decision-making by statistical modelling and machine learning. The MS-TDS predicts the 24-month outcome of no new or enlarging cMRI lesions in newly diagnosed CIS and RRMS patients. It provides risk estimates that can be used to identify patients who are expected to benefit from no medication or platform medication. The overall performance of the prediction model is weak but comparable to similar models that have recently been suggested. A prospective study is currently being conducted to allow for external validation. The clinical relevance of MS-TDS has yet to be demonstrated. The task of developing models for supporting treatment decisions in early MS remains challenging, and the present work can serve as a methodological example for future studies.

Declarations

Ethics approval and consent to participate

The relevant Ethics Committee of the Technical University of Munich approved the study (approval number: 21/20 S-KH) and waived the requirement for informed consent to participate according to Article 27 of the Bavarian Hospital Act. Patient data were used for research purposes only. For the planned secondary use of already existing routine data, no further information and consent of the patients was necessary.

Consent for publication

Individual person data are not provided, making informed consent for publication unnecessary.

Author contributions

Alexander Hapfelmeier: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Begum Irmak On: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review & editing.

Mark Mühlau: Conceptualization; Data curation; Writing – original draft; Writing – review & editing.

Jan S. Kirschke: Conceptualization; Data curation; Resources; Writing – original draft; Writing – review & editing.

Achim Berthele: Conceptualization; Data curation; Writing – review & editing.

Christiane Gasperi: Conceptualization; Data curation; Writing – original draft; Writing – review & editing.

Ulrich Mansmann: Conceptualization; Funding acquisition; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.

Alexander Wuschek: Conceptualization; Data curation; Writing – review & editing.

Matthias Bussas: Conceptualization; Data curation; Writing – review & editing.

Martin Boeker: Conceptualization; Data curation; Project administration; Resources; Writing – original draft; Writing – review & editing.

Antonios Bayas: Conceptualization; Writing – review & editing.

Makbule Senel: Conceptualization; Writing – review & editing.

Joachim Havla: Conceptualization; Writing – review & editing.

Markus C. Kowarik: Conceptualization; Writing – review & editing.

Klaus Kuhn: Conceptualization; Funding acquisition; Writing – review & editing.

Ingrid Gatz: Conceptualization; Data curation; Writing – review & editing.

Helmut Spengler: Conceptualization; Data curation; Writing – review & editing.

Benedikt Wiestler: Conceptualization; Data curation; Writing – review & editing.

Lioba Grundl: Conceptualization; Data curation; Writing – review & editing.

Dominik Sepp: Conceptualization; Data curation; Writing – review & editing.

Bernhard Hemmer: Conceptualization; Data curation; Funding acquisition; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.

Acknowledgements

The authors thank Nikolaus Will, Marie-Christin Metz, David Schinz, Dominik Heim, Philip Prucker, Benita Schnitz-Koep and Daria Filatova for the image segmentation.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The project was funded by the German Federal Ministry of Education and Research (funding code 01ZZ1804A and 01ZZ1804C). The responsibility for the content of this publication lies with the authors.

Competing interests

The authors declared the following potential conflicts of interest with respect to the research, authorship and/or publication of this article: A.H., B.I.O., M.M., A.Be., U.M., A.W., M.Bu., M.Bo., K.K., I.G., H.S., B.W., L.G. and D.S. declare that there is no conflict of interest. J.S.K. is Co-Founder of Bonescreen GmbH. C.G. reports funding from the German Research Foundation (Deutsche Forschungsgesellschaft DFG), the Hertie Foundation, the Hans and Klementia Langmatz and the German Federal Ministry of Education and Research, all of which are not related to this study. A.Ba. reports personal compensation from Merck Serono, Biogen, Novartis, TEVA, Roche, Sanofi/Genzyme, Celgene/Bristol Myers Squibb, Janssen, Sandoz/HEXAL; grants for congress travel and participation from Biogen, TEVA, Novartis, Sanofi/Genzyme, Merck Serono, Celgene and Janssen. None related to this report. M.S. has received consulting and/or speaker honoraria from Alexion, Bayer, Biogen, Bristol Myers Squibb, Merck, Roche and Sanofi Genzyme. She has received travel support from Celgene and TEVA. She has received research funding from the Hertha-Nathorff-Program. None of this related to this study. M.C.K. has served on advisory boards and received speaker fees/travel grants from Merck, Sanofi/Genzyme, Novartis, Biogen, Jansen, Alexion, Celgene/Bristol Myers Squibb and Roche. M.K. also received research grants from Merck, Sanofi/Genzyme and Celgene/Bristol Myers Squibb, Novartis and Janssen, all not related to this study. J.H. reports grants for OCT research from the Friedrich-Baur-Stiftung and Merck, personal fees and nonfinancial support from Celgene, Horizon, Janssen, Bayer, Merck, Alexion, Novartis, Roche, Biogen and non-financial support of the Guthy-Jackson Charitable Foundation, all outside the submitted work. J.H. is partially funded by the German Federal Ministry of Education and Research [(DIFUTURE), grant numbers 01ZZ1603[A-D] and 01ZZ1804[A-H]]. B.H. has served on scientific advisory boards for Novartis; he has served as DMSC member for AllergyCare, Polpharma, Sandoz and TG therapeutics; he or his institution have received speaker honoraria from Desitin; his institution received research grants from Regeneron for multiple sclerosis research. B.H. holds part of two patents; one for the detection of antibodies against KIR4.1 in a subpopulation of patients with multiple sclerosis and one for genetic determinants of neutralizing antibodies to interferon. All of B.H.'s conflicts are not relevant to the topic of the study. B.H. received funding from the Multiple MS EU consortium, the Clinspect-M consortium funded by the Bundesministerium für Bildung und Forschung and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the framework of the Munich Cluster for Systems Neurology (EXC 2145 SyNergy – ID 390857198).

Availability of data and materials

Individual patient data cannot be shared with third parties due to legal restrictions. Analysis codes can be provided upon justified request.

ORCID iDs

Alexander Hapfelmeier https://orcid.org/0000-0001-6765-6352

Ulrich Mansmann https://orcid.org/0000-0002-9955-8906

Makbule Senel https://orcid.org/0000-0002-2737-7495

Joachim Havla https://orcid.org/0000-0002-4386-1340

Markus C. Kowarik https://orcid.org/0000-0003-1389-5539

References

1. Daltrozzo T, Hapfelmeier A, Donnachie E, *et al.* A systematic assessment of prevalence, incidence and regional distribution of multiple sclerosis in Bavaria from 2006 to 2015. *Front Neurol* 2018; 9: 871.
2. Walton C, King R, Rechtman L, *et al.* Rising prevalence of multiple sclerosis worldwide: insights from the Atlas of MS, third edition. *Mult Scler* 2020; 26: 1816–1821.
3. Thompson AJ, Banwell BL, Barkhof F, *et al.* Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol* 2018; 17: 162–173.
4. Chung KK, Altmann D, Barkhof F, *et al.* A 30-year clinical and magnetic resonance imaging observational study of multiple sclerosis and clinically isolated syndromes. *Ann Neurol* 2020; 87: 63–74.
5. Koch-Henriksen N and Sørensen PS. The changing demographic pattern of multiple sclerosis epidemiology. *Lancet Neurol* 2010; 9: 520–532.
6. Tintore M, Arrambide G, Otero-Romero S, *et al.* The long-term outcomes of CIS patients in the Barcelona inception cohort: looking back to recognize aggressive MS. *Mult Scler* 2020; 26: 1658–1669.
7. Brown FS, Glasmacher SA, Kearns PKA, *et al.* Systematic review of prediction models in relapsing remitting multiple sclerosis. *PLoS ONE* 2020; 15: e0233575.
8. Havas J, Leray E, Rollot F, *et al.* Predictive medicine in multiple sclerosis: a systematic review. *Mult Scler Relat Disord* 2020; 40: 101928.
9. Seccia R, Romano S, Salvetti M, *et al.* Machine learning use for prognostic purposes in multiple sclerosis. *Life* 2021; 11: 122.
10. Pellegrini F, Copetti M, Sormani MP, *et al.* Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. *Mult Scler* 2020; 26: 1828–1836.
11. Chalkou K, Steyerberg E, Bossuyt P, *et al.* Development, validation and clinical usefulness of a prognostic model for relapse in relapsing-remitting multiple sclerosis. *Diagn Progn Res* 2021; 5: 17.
12. De Brouwer E, Becker T, Moreau Y, *et al.* Corrigendum to Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression: [Computer Methods and Programs in Biomedicine, Volume 208, (September 2021) 106180]. *Comput Methods Programs Biomed* 2022; 213: 106479.
13. Jacobs LD, Beck RW, Simon JH, *et al.* Intramuscular interferon beta-1a therapy initiated during a first demyelinating event in multiple sclerosis. CHAMPS Study Group. *N Engl J Med* 2000; 343: 898–904.
14. Rudick RA, Stuart WH, Calabresi PA, *et al.* Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med* 2006; 354: 911–923.
15. Kappos L, Bar-Or A, Cree BAC, *et al.* Siponimod versus placebo in secondary progressive multiple sclerosis (EXPAND): a double-blind, randomised, phase 3 study. *Lancet* 2018; 391: 1263–1273.
16. Hauser SL, Bar-Or A, Cohen JA, *et al.* Ofatumumab versus teriflunomide in multiple sclerosis. *N Engl J Med* 2020; 383: 546–557.
17. van Munster CE and Uitdehaag BM. Outcome measures in clinical trials for multiple sclerosis. *CNS Drugs* 2017; 31: 217–236.
18. Bayas A, Hoffmann VS, Berthele A, *et al.* A new multi-centre prospective cohort of patients with clinically isolated syndrome and early relapsing-remitting multiple sclerosis: rationale and baseline features of ProVal-MS. *Eur J Pharm Med Res*, submitted.
19. Wattjes MP, Rovira À, Miller D, *et al.* Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple

sclerosis – establishing disease prognosis and monitoring patients. *Nat Rev Neurol* 2015; 11: 597–606.

20. Schmidt P, Gaser C, Arsic M, *et al*. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *NeuroImage* 2012; 59: 3774–3783.
21. Isensee F, Jaeger PF, Kohl SAA, *et al*. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 2021; 18: 203–211.
22. Yushkevich PA, Piven J, Hazlett HC, *et al*. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *NeuroImage* 2006; 31: 1116–1128.
23. Eichinger P, Schön S, Pongratz V, *et al*. Accuracy of unenhanced MRI in the detection of new brain lesions in multiple sclerosis. *Radiology* 2019; 291: 429–435.
24. Hothorn T, Möst L and Bühlmann P. Most likely transformations. *Scand J Stat* 2018; 45: 110–134.
25. Korepanova N, Seibold H, Steffen V, *et al*. Survival forests under test: impact of the proportional hazards assumption on prognostic and predictive forests for amyotrophic lateral sclerosis survival. *Stat Methods Med Res* 2020; 29: 1403–1419.
26. Liaw A and Wiener M. Classification and regression by randomForest. *R News* 2002; 2: 18–22.
27. Hothorn T. trtf: transformation trees and forests (R package version 0.3-7 ed.), 2020.
28. Hothorn T. Most likely transformations: the mlt package. *J Stat Softw* 2020; 92: 1–68.
29. Díaz-Coto S, Martínez-Camblor P and Corral-Blanco NO. Cumulative/dynamic ROC curve estimation under interval censorship. *J Stat Comput Simul* 2020; 90: 1570–1590.
30. Hothorn T and Zeileis A. Predictive distribution modeling using transformation forests. *J Comput Graph Stat* 2021; 30: 1181–1196.
31. Hapfelmeier A, Hornung R and Haller B. Efficient permutation testing of variable importance measures by the example of random forests. *Comput Stat Data Anal* 2023; 181: 107689.
32. Strobl C, Malley J and Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 2009; 14: 323–348.
33. Stekhoven DJ and Bühlmann P. MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics* 2012; 28: 112–118.
34. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg* 2015; 102: 148–158.
35. Trojano M, Tintore M, Montalban X, *et al*. Treatment decisions in multiple sclerosis – insights from real-world observational studies. *Nat Rev Neurol* 2017; 13: 105–118.
36. Lu M, Sadiq S, Feaster DJ, *et al*. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat* 2018; 27: 209–219.
37. Powers S, Qian J, Jung K, *et al*. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med* 2018; 37: 1767–1787.
38. Dandl S, Hothorn T, Seibold H, *et al*. What makes forest-based heterogeneous treatment effect estimators work? ArXiv 2022; preprint arXiv:220610323.
39. Confavreux C, Vukusic S and Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. *Brain* 2003; 126: 770–782.
40. Fisniku LK, Brex PA, Altmann DR, *et al*. Disability and T2 MRI lesions: a 20-year follow-up of patients with relapse onset of multiple sclerosis. *Brain* 2008; 131: 808–817.
41. Gajofatto A, Calabrese M, Benedetti MD, *et al*. Clinical, MRI, and CSF markers of disability progression in multiple sclerosis. *Dis Markers* 2013; 35: 687–699.
42. Stühler E, Braune S, Lionetto F, *et al*. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Med Res Methodol* 2020; 20: 24.
43. Hosmer DW Jr, Lemeshow S and Sturdivant RX. *Applied Logistic Regression*. Hoboken, NJ: John Wiley, 2013.

Appendix 1. Timeframes and definitions of patient features.

Numeric variables:				
Domain/test	**Variable name**	**Type**	**Unit**	**Timespan**
BDI	Test result (depression)	Integer		The earliest of closest to baseline during ±3 months
FSMC	Fatigue cognitive	Integer		The earliest of closest to baseline during ±3 months
FSMC	Fatigue motor	Integer		The earliest of closest to baseline during ±3 months
FSMC	Total	Integer		The earliest of closest to baseline during ±3 months
MSFC	Nine-Hole Peg test result – hand/arm	Numeric		The earliest of closest to baseline during ±3 months
MSFC	25-Foot Walk test result – ambulation	Numeric		The earliest of closest to baseline during ±3 months
CSF	Leucocyte count	Integer	cnt/mcl	The earliest of closest to baseline during ±3 months
CSF	Glucose	Integer	mg/dl	The earliest of closest to baseline during ±3 months
CSF	Total protein	Integer	mg/dl	The earliest of closest to baseline during ±3 months
CSF	Albumin quotient	Numeric	$\times 10^{-3}$	The earliest of closest to baseline during ±3 months
CSF	CSF-specific oligoclonal bands	Ordered	no/borderline/yes	The earliest of closest to baseline during ±3 months
CSF	IgG Quo/Alb Quo	Numeric		The earliest of closest to baseline during ±3 months
CSF	IgM Quo/Alb Quo	Numeric		The earliest of closest to baseline during ±3 months
CSF	IgA Quo/Alb Quo	Numeric		The earliest of closest to baseline during ±3 months
EDSS	Total	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Pyramidal	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Cerebellar	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Brainstem	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Sensory	Ordered		The earliest of closest to baseline during ±3 months

(Continued)

Appendix 1. (Continued)

Domain/test	Variable name	Type	Unit	Timespan
EDSS	Bowel and bladder	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Visual	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Cognitive	Ordered		The earliest of closest to baseline during ±3 months
EDSS	Ambulation	Ordered		The earliest of closest to baseline during ±3 months
Laboratory	Basophils	Numeric	%	The earliest of closest to baseline during ±3 months
Laboratory	Bilirubin	Numeric	mg/dl	The earliest of closest to baseline during ±3 months
Laboratory	Blood urea nitrogen	Numeric	mg/dl	The earliest of closest to baseline during ±3 months
Laboratory	Eosinophils	Numeric	%	The earliest of closest to baseline during ±3 months
Laboratory	Erythrocytes	Numeric	cnt/pl	The earliest of closest to baseline during ±3 months
Laboratory	GOT ASAT	Numeric	U/l	The earliest of closest to baseline during ±3 months
Laboratory	GPT ALAT	Numeric	U/l	The earliest of closest to baseline during ±3 months
Laboratory	Haematocrit	Numeric	%	The earliest of closest to baseline during ±3 months
Laboratory	Haemoglobin	Numeric	g/dl	The earliest of closest to baseline during ±3 months
Laboratory	Leucocytes	Numeric	cnt/nl	The earliest of closest to baseline during ±3 months
Laboratory	Lymphocytes	Numeric	%	The earliest of closest to baseline during ±3 months
Laboratory	MCH	Numeric	pg	The earliest of closest to baseline during ±3 months
Laboratory	MCHC	Numeric	g/dl	The earliest of closest to baseline during ±3 months
Laboratory	MCV	Numeric	fl	The earliest of closest to baseline during ±3 months
Laboratory	Monocytes	Numeric	%	The earliest of closest to baseline during ±3 months
Laboratory	Neutrophils	Numeric	%	The earliest of closest to baseline during ±3 months

(Continued)

Appendix 1. (Continued)

Domain/test	Variable name	Type	Unit	Timespan
Laboratory	Thrombocytes	Numeric	cnt/nl	The earliest of closest to baseline during ±3 months
Laboratory	TSH	Numeric	mcIU/ml	The earliest of closest to baseline during ±3 months
Demographics	Age	Numeric	Years	N/A
Relapses	Number of relapses	Numeric	Count	During 3 years before baseline
Demographics	Height	Numeric	cm	The earliest of closest to ever before than +3 months from baseline
Demographics	Weight	Numeric	kg	The earliest of closest around baseline
Demographics	BMI	Numeric	kg/m^2	The earliest of closest to ever before than +3 months from baseline
Diagnosis	Time since diagnosis at baseline	Integer	Days	The earliest of closest to baseline during –36/+1 months
cMRI	Total number of spatially separated lesions	Integer	Enhancing and non-enhancing	First image

ASAT, aspartate transaminase; ALAT, alanine transaminase; BDI, Beck Depression Inventory; BMI, body mass index; CIS, clinically isolated syndrome; cMRI, cerebral magnetic resonance images; CSF, Cerebrospinal fluid; EDSS, Expanded Disability Severity Scale; FSMC, Fatigue Scale for Motor and Cognitive Functions; GOT, glutamic oxaloacetic transaminase; GPT, glutamic-pyruvic transaminase; IQR, interquartile range; MCH, mean corpuscular haemoglobin; MCHC, mean corpuscular haemoglobin concentration; MCV, mean corpuscular volume; MSFC, MS functional composite; N/A, not applicable; Quo, quotient; TSH, thyroid-stimulating hormone.

Categorical variables:				
Domain/test	**Variable name**	**Type**	**Categories**	**Timespan**
Demographics	Smoking	Ordered	Yes/former/no/unknown	Ever smoker (at base document or following visits) ELSE/ever ex-smoker (at base document or following visits) ELSE/ever nonsmoker (at base document or following visits) ELSE/no information at base document or following visits
First symptom[a,b]	Numbness	Logical	No/yes	Ever before baseline
First symptom[a,b]	Other cranial nerve symptom	Logical	No/yes	Ever before baseline

(Continued)

Appendix 1. (Continued)

Domain/test	Variable name	Type	Categories	Timespan
First symptom[a,b]	Paresis	Logical	No/yes	Ever before baseline
First symptom[a,b]	Optic neuritis	Logical	No/yes	Ever before baseline
First symptom[a,b]	Any other symptom	Logical	No/yes	Ever before baseline
Demographics	Sex	Factor	Male/female	N/A
Relapses[a,b]	Numbness	Logical	No/yes	During ±3 months from baseline
Relapses[a,b]	Other neurological symptom	Logical	No/yes	During ±3 months from baseline
Relapses[a,b]	Paresis	Logical	No/yes	During ±3 months from baseline
Relapses[a,b]	Optic neuritis	Logical	No/yes	During ±3 months from baseline
Relapses[a,b]	Any other symptom	Logical	No/yes	During ±3 months from baseline
cMRI	Periventricular lesions	Logical	No/yes	First image
cMRI	Subcortical/ unspecific lesions	Logical	No/yes	First image
cMRI	Juxtacortical or cortical lesions	Logical	No/yes	First image
cMRI	Infratentorial lesions	Logical	No/yes	First image
Diagnosis	Diagnosis at baseline (CIS)	Factor	CIS/RRMS	Earliest of closest to baseline during −36/+1 months

cMRI, cerebral magnetic resonance images; CIS, clinically isolated syndrome; RRMS, relapsing-remitting multiple sclerosis.
[a]No if no information.
[b]Include if only year is available in the same year as baseline.

Appendix 2

Session Info

```
> sessionInfo()
R version 3.6.3 (2020-02-29)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.6 LTS

Matrix products: default
BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
```

```
locale:
[1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C       LC_TIME=C.UTF-8
[4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
[7] LC_PAPER=C.UTF-8       LC_NAME=C      LC_ADDRESS=C
[10] LC_TELEPHONE=C        LC_MEASUREMENT=C.UTF-8     LC_IDENTIFICATION=C

attached base packages:
[1] parallel grid      stats      graphics grDevices utils      datasets
[8] methods base

other attached packages:
[1] profvis_0.3.7      tictoc_1.0.1       icenReg_2.0.15
[4] coda_0.19-4        Rcpp_1.0.7         intcensROC_0.1.3
[7] ALassoSurvIC_0.1.0     caret_6.0-90       lattice_0.20-45
[10] trtf_0.3-8        partykit_1.2-15        mvtnorm_1.1-3
[13] libcoin_1.0-9         tram_0.6-1         mlt_1.3-2
[16] basefun_1.1-1         variables_1.1-1        missForest_1.4
[19] itertools_0.1-3       iterators_1.0.13       foreach_1.5.1
[22] randomForest_4.6-14       survival_3.2-13        BBmisc_1.11
[25] summarytools_1.0.0        forcats_0.5.1      stringr_1.4.0
[28] dplyr_1.0.7       purrr_0.3.4        readr_2.1.1
[31] tidyr_1.1.4       tibble_3.1.6       ggplot2_3.3.5
[34] tidyverse_1.3.1       openxlsx_4.2.4         xtable_1.8-4
[37] rio_0.5.29        knitr_1.36         reshape2_1.4.4
[40] lubridate_1.8.0       readxl_1.3.1

loaded via a namespace (and not attached):
[1] TH.data_1.1-0      colorspace_2.0-2       pryr_0.1.5
[4] class_7.3-19       ellipsis_0.3.2         base64enc_0.1-3
[7] fs_1.5.1       rstudioapi_0.13        listenv_0.8.0
[10] prodlim_2019.11.13        fansi_0.5.0        xml2_1.3.3
[13] codetools_0.2-18      splines_3.6.3      polynom_1.4-0
[16] Formula_1.2-4         jsonlite_1.7.2         nloptr_1.2.2.3
[19] pROC_1.18.0       broom_0.7.10       dbplyr_2.1.1
[22] compiler_3.6.3        httr_1.4.2         backports_1.4.0
[25] assertthat_0.2.1      Matrix_1.3-4       fastmap_1.1.0
[28] cli_3.1.0         htmltools_0.5.2        tools_3.6.3
[31] gtable_0.3.0      glue_1.5.1         cellranger_1.1.0
[34] vctrs_0.3.8       nlme_3.1-153       timeDate_3043.102
[37] inum_1.0-4        gower_0.2.2        xfun_0.28
[40] globals_0.14.0        rvest_1.0.2        lifecycle_1.0.1
[43] future_1.23.0         MASS_7.3-54        zoo_1.8-9
[46] scales_1.1.1      ipred_0.9-12       hms_1.1.1
[49] sandwich_3.0-1        curl_4.3.2         pander_0.6.4
[52] rpart_4.1-15      stringi_1.7.6      checkmate_2.0.0
[55] orthopolynom_1.0-5        BB_2019.10-1       zip_2.2.0
[58] lava_1.6.10       rlang_0.4.12       pkgconfig_2.0.3
[61] matrixStats_0.61.0        htmlwidgets_1.5.4      recipes_0.1.17
[64] rapportools_1.0       tidyselect_1.1.1       parallelly_1.29.0
[67] plyr_1.8.6        magrittr_2.0.1         R6_2.5.1
[70] magick_2.7.3      generics_0.1.1         multcomp_1.4-17
[73] DBI_1.1.1         pillar_1.6.4       haven_2.4.3
```

[76] foreign_0.8-76 withr_2.4.3 nnet_7.3-16
[79] future.apply_1.8.1 modelr_0.1.8 crayon_1.4.2
[82] coneproj_1.14 utf8_1.2.2 alabama_2015.3-1
[85] tzdb_0.2.0 data.table_1.14.2 ModelMetrics_1.2.2.2
[88] reprex_2.0.1 digest_0.6.29 numDeriv_2016.8-1.1
[91] stats4_3.6.3 munsell_0.5.0 tcltk_3.6.3
[94] quadprog_1.5-8

Appendix 3

Detailed results of the benchmark study

In the three inner loops of the benchmark study conducted for model selection and performance estimation, one of the selected models used the parameters minsplit = 40 and mtry = $p/3 = 22$ and two of the selected models each used the parameters minsplit = 30 and mtry = $p/3 = 22$. Average area under the receiver operating characteristics curves (AUROCs) of these models were 0.598, 0.587 and 0.617 in the inner loops. The AUROCs obtained on the respective test data of the outer loop were 0.676, 0.588 and 0.607, leading to an overall estimate of the average performance of a best model of 0.624. In the outer loop, the parameter settings minsplit = 20 and mtry = $p/3 = 22$ performed best and were therefore used in refitting the final model to the whole data set.

Appendix 4. TRIPOD checklist for prediction model development and validation.

Section/topic	Item[a]		Checklist item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and validating a multivariable prediction model, the target population and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results and conclusions.	1
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	2
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	2
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g. randomized trial, cohort or registry data), separately for the development and validation data sets, if applicable.	3–4
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	3

(Continued)

Appendix 4. (Continued)

Section/topic	Item[a]		Checklist item	Page
Participants	5a	D;V	Specify key elements of the study setting (e.g. primary care, secondary care, general population) including number and location of centres.	3
	5b	D;V	Describe eligibility criteria for participants.	3
	5c	D;V	Give details of treatments received, if relevant.	3
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	4
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	3–4, Table 1
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	3
Missing data	9	D;V	Describe how missing data were handled (e.g. complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	5
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	5
	10b	D	Specify type of model, all model-building procedures (including any predictor selection) and method for internal validation.	5
	10c	V	For validation, describe how the predictions were calculated.	5
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	5
	10e	V	Describe any model updating (e.g. recalibration) arising from the validation, if done.	5
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development *versus* validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome and predictors.	NA
Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	5–6, Figure 2

(Continued)

Appendix 4. (Continued)

Section/topic	Item[a]		Checklist item	Page
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	6, Table 1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	6
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e. all regression coefficients and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to the use the prediction model.	10–11
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	6
Model-updating	17	V	If done, report the results from any model updating (i.e. model specification, model performance).	6
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	13
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	11
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies and other relevant evidence.	13–14
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	14
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator and data sets.	18–23
Funding	22	D;V	Give the source of funding and the role of the funders for this study.	15

TRIPOD, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis; NA: not applicable.
[a]Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V and items relating to both are denoted D;V.