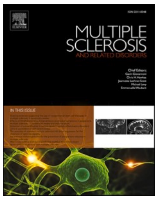
Multiple Sclerosis and Related Disorders 93 (2025) 106247

Contents lists available at ScienceDirect

Multiple Sclerosis and Related Disorders

journal homepage: www.elsevier.com/locate/msard



Multivariable prognostic prediction of efficacy and safety outcomes and response to fingolimod in people with relapsing-remitting multiple sclerosis

Begüm Irmak Ön [a,b,*], Joachim Havla [c], Ulrich Mansmann [a,b]

[a] Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, LMU Munich, Marchioninistrasse 15 81377, Munich, Germany
[b] Pettenkofer School of Public Health, Elisabeth-Winterhalter-Weg 6, 81377 Munich, Germany
[c] Institute of Clinical Neuroimmunology, University Hospital, LMU Munich, Marchioninistrasse 15 81377, Munich, Germany

ARTICLE INFO

Keywords:
Relapsing-Remitting Multiple Sclerosis
Fingolimod
Prognosis
Individualized Medicine
Treatment Effect Heterogeneity
Clinical Prediction Rule

ABSTRACT

Background: The individual treatment response in people with relapsing-remitting multiple sclerosis (RRMS) remain unpredictable. In order to support medical decisions, we aimed to predict response to fingolimod compared to placebo, by developing and validating prognostic multivariable models.

Methods: We included two-year follow-up from intention-to-treat populations of two multi-country placebo-controlled randomized controlled trials (RCT) of daily fingolimod 0.5 mg. The data was accessed via ClinicalStudyDataRequest.com (Proposal Number: 11223) The RCTs were in adult RRMS patients with active disease. We used four Cox proportional hazards based penalized (elastic net and grouped lasso) and tree methods (transformation tree and forest) to predict time-to relapse and other relevant efficacy and safety endpoints in data from the RCT FREEDOMS. Treatment arm, 80 baseline variables and their interaction with treatment were considered as candidate predictors in the models. A nested cross-validation scheme ensured independent tuning parameter optimization and internal model performance evaluation. The generalizability of the models with the highest cross-validated time-dependent area under the receiver operating curve (AUC) was further evaluated in terms of discrimination (AUC), calibration (plots, intercept, slope), clinical utility (decision curve analysis), and treatment response plots by external validation in data from the RCT FREEDOMS II.

Results: The best performing model predicting relapse risk (331 events) in the development sample (n=843) was an elastic net regression with main terms for four predictors alongside treatment: EDSS score, volume of Gadolinium enhanced T1 lesions, number of relapses in the last 2 years, and number of prior MS treatments. In external validation (n=713), it had an AUC of 0.68 (95% CI 0.63–0.72), but the predictions were overestimating the actual risk (358 events) with a calibration-in-the-large of -0.17 (-0.3 - -0.04) and a slope of 1.06 (0.78–1.35). Almost no heterogeneity (variability 0.001) was detected in the predicted relapse risk change in response to fingolimod. FREEDOMS II participants were predicted to have 0.21 to 0.31 absolute relapse risk reduction with fingolimod compared to placebo. The selected model predicting new or enlarging T2 magnetic resonance imaging (MRI) lesions had an AUC of 0.74 (0.70–0.78), moderate calibration, but no treatment response variability. The final model predicting confirmed disability progression had an AUC of 0.59 (0.54–0.64) and the predicted treatment response heterogeneity was not significant. The overall safety outcome could not be predicted with sufficient discrimination. However, the final model predicting infections or neoplasms had an AUC of 0.69 (0.63–0.74) and non-significant treatment response heterogeneity. For the efficacy outcomes, important predictors were related to (para)clinical disease activity or disability. Unexpected influential predictors included concomitant disorders.

Conclusion: Relapse and new or enlarging T2 MRI lesions were moderately predictable in an independent sample with the developed prognostic models. Fingolimod was expected to decrease the risk of these events for all patients, with no predictable heterogeneity. Disability and safety outcomes could not be well-predicted and it is yet unresolved whether the change in their risk as response to fingolimod is heterogeneous or not.

* Correspondence author at: Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Marchioninistrasse 15, 81377 Munich, Germany.
E-mail addresses: ionseker@ibe.med.uni-muenchen.de (B. Irmak Ön), Joachim.Havla@med.uni-muenchen.de (J. Havla), mansmann@ibe.med.uni-muenchen.de (U. Mansmann).

https://doi.org/10.1016/j.msard.2024.106247
Received 26 August 2024; Received in revised form 5 November 2024; Accepted 20 December 2024
Available online 21 December 2024

Abbreviations

AUC	area under the receiver operating curve
AUC(t)	time-dependent area under the curve
Brier(t)	time-dependent Brier score
DCA	decision curve analysis
EPV	events per variable
FREEDOMS	FTY720 Research Evaluating Effects of Daily Oral therapy in Multiple Sclerosis
ROC	receiver operating characteristic
SI	supplementary information
TRIPOD+AI	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + Artificial Intellegence
3m-CDP	disability progression confirmed 3 months after its onset
9HPT	9-hole peg test
AE	adverse event
CI	confidence interval
DMT	dozen disease-modifying treatments
EDSS	expanded disability status scale
Gd	gadolinium
KM	Kaplan-Meier
MRI	magnetic resonance imaging
PH	proportional hazards
RCT	randomized controlled trial
RRMS	relapsing-remitting multiple sclerosis
SAE	serious adverse event
T2	T2-weighted

1. Introduction

Relapsing-Remitting Multiple Sclerosis (RRMS) is a chronic debilitating disease. The progress and severity of the disease is described as highly heterogeneous (Ziemssen et al., 2019) and difficult to predict (Brown et al., 2020; Havas et al., 2020). The availability of more than a dozen disease-modifying treatments (DMT), the individual responses to which are also considered to vary, compounds the difficulty of clinical decision making. Prior identification of patients likely to benefit or be harmed from a specific treatment can be useful (Baecher-Allan et al., 2018). Subgroup analyses identify differential treatment response for a limited number of binary-coded patient characteristics, with drawbacks like reference class, inflated error rates, lack of sufficient adjustment, and neglecting safety (Wang et al., 2007). Prognostic models predicting individual health outcomes or treatment response constitute an alternative to relying on average event rates or rough groups (Kent et al., 2018).

The bulk of the published prognostic models in RRMS patients lack external validations, are at risk of overoptimistically-biased results, have difficult-to-collect predictors and do not comply with the reporting guidelines (Reeve et al., 2023). Their widespread implementation is hindered by the lack of methodological compliance for valid and accessible models (Steyerberg et al., 2013). Additionally, it is unclear if there is predictable treatment effect heterogeneity or a prognostic model would suffice to explain the observed heterogeneity. Multivariable treatment response prediction in the ideal setting of randomized clinical trial (RCT), was employed for a few DMTs and efficacy outcomes (Chalkou et al., 2021b; Pellegrini et al., 2020a) with mixed results.

Fingolimod is a high-efficacy treatment option for RRMS patients used as first- or second-line treatment in RRMS patients (Wiendl et al., 2021). The immunosuppressant nature of DMTs may induce leukopenia or lymphopenia, increasing infections and neoplasms (Winkelmann et al., 2016). A published prediction of individual response to fingolimod is missing, although it is an increasingly utilized DMT (Holstiege et al., 2022). Studies in Germany identified fingolimod to be most frequently used after first-line DMTs, received by 10–14% of RRMS patients (Müller et al., 2020; Ohlmeier et al., 2020). Around 18–30% of RRMS patients did not receive any DMT, representing the option of no treatment.

To personalize medicine in RRMS, we developed and externally validated multivariable models predicting prognosis and response to fingolimod. The primary endpoint was relapse within 2 years. Other efficacy and safety endpoints were also predicted. Finally, we identified variables predictive of these endpoints.

2. Material and methods

We performed the analysis in *R* (4.2.0), see Supplementary Information (SI) SI1. For reporting, we used TRIPOD+AI (SI2).

2.1. Design

The datasets from the multi-country phase III FREEDOMS (01.2006–08.2007, 22 countries) and FREEDOMS II (06.2006–03.2009, eight countries) (ClinicalTrials.gov NCT00289978 and NCT00355134) RCTs were repurposed and used respectively in model development and external validation (Fig. 1). The primary objective of both trials was to compare the 24-month relapse rate in RRMS patients randomized to daily placebo, fingolimod 0.5 mg, or fingolimod 1.25 mg (1:1:1). Visits and examinations occurred at baseline, two weeks, one, two, three months, and every three months afterwards. MRI scans were taken at baseline and months six, 12, and 24.

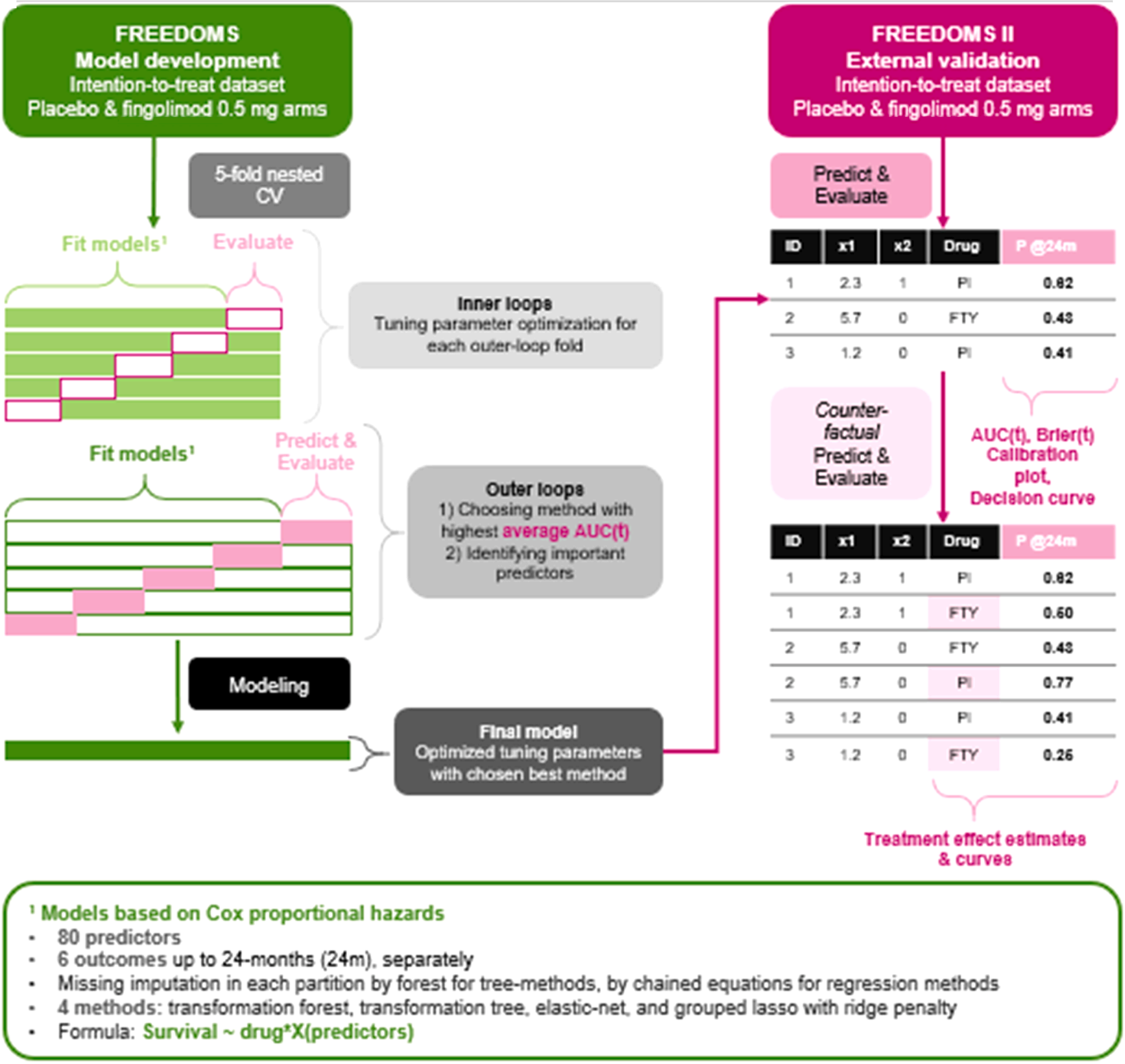
Approvals from institutional review boards and patient informed consents were in place for the FREEDOMS trials and their results are reported elsewhere(Calabresi et al., 2014; Derfuss et al., 2016; Devonshire et al., 2012; Kappos et al., 2010). The anonymized data were made available to us by Novartis via the data sharing platform www.clinicalstudydatarequest.com (Proposal Number: 11223) until end of 2023. This project was deemed exempt from approval by the Ethics Committee of LMU Munich (Project Number: 19–838).

FREEDOMS trials included RRMS patients aged 18–55 years, and diagnosed based on the McDonald 2005 criteria. The participants had to have an EDSS score lower than 6.0, and at least one relapse during the year or two relapses during the two years before randomization. A time gap between a previous DMT and randomization was also required. In our study, we included the intention-to-treat population randomized to the approved daily dose of fingolimod, 0.5 mg, or placebo, representing the option of no treatment.

2.2. Variables

In addition to the randomized drug as placebo or fingolimod 0.5 mg, we considered 80 potentially prognostic variables at baseline as candidates, covering various domains like demographic and clinical parameters, comedications, concomitant diseases, laboratory and quality of life measures (SI3 Table S1). Except from age (eight categories), all categorical variables were binary. The modeling methods implicitly or explicitly included interaction terms of drug with all the candidate predictors, bringing the number of terms considered to 175.

The primary outcome was confirmed relapse, as defined in the trials. The other efficacy outcomes were new or enlarging lesions in T2 MRI scans (T2 MRI), and disability progression confirmed 3 months after its onset (3m-CDP). We investigated a general safety outcome of any serious adverse event (SAE) or trial discontinuation due to an adverse event (AE) (Safety), and an AE from the system organ classes of infections and infestations or neoplasms (Immune safety). Parameters were standardized and assessed by trained investigators blinded to the treatment assignment. The endpoints were time-to-first event since randomization until the 24-month visit (720 days). For stability, we censored observations without event on the first of the participant's last visit or day 765.


FREEDOMS
Model development
Intention-to-treat dataset
Placebo & fingolimod 0.5 mg arms
5-fold nested CV
Fit models[1]
Evaluate
Inner loops
Tuning parameter optimization for each outer-loop fold
Fit models[1]
Predict & Evaluate
Outer loops
1) Choosing method with highest average AUC(t)
2) Identifying important predictors
Modeling
Final model
Optimized tuning parameters with chosen best method
FREEDOMS II
External validation
Intention-to-treat dataset
Placebo & fingolimod 0.5 mg arms
Predict & Evaluate
Counter-factual Predict & Evaluate
AUC(t), Brier(t)
Calibration plot,
Decision curve
Treatment effect estimates & curves

ID	x1	x2	Drug	P @24m
1	2.3	1	Pl	0.82
2	5.7	0	FTY	0.48
3	1.2	0	Pl	0.41

ID	x1	x2	Drug	P @24m
1	2.3	1	Pl	0.82
1	2.3	1	FTY	0.60
2	5.7	0	FTY	0.48
2	5.7	0	Pl	0.77
3	1.2	0	Pl	0.41
3	1.2	0	FTY	0.26

[1] Models based on Cox proportional hazards
- 80 predictors
- 6 outcomes up to 24-months (24m), separately
- Missing imputation in each partition by forest for tree-methods, by chained equations for regression methods
- 4 methods: transformation forest, transformation tree, elastic-net, and grouped lasso with ridge penalty
- Formula: Survival ~ drug*X(predictors)

Fig. 1. Overview of methods
Methods employed to develop and externally validate prognostic and treatment response prediction models in the two randomized controlled trial datasets: FREEDOMS and FREEDOMS II. CV: cross-validation, AUC(t): time-dependent area under the receiver operator characteristic curve, P@24m: risk of event at 24-months, FTY: fingolimod 0.5 mg, Pl: placebo.

2.3. *Statistical methods*

We described the treatment arms in the development and validation samples using median and range or by frequencies. We reported event numbers and stratified Kaplan-Meier (KM) curves. For sample size considerations, we report events per variable (EPV) in the development sample and the event numbers in the validation (Moons et al., 2019).

2.3.1. *Development*

We considered four modeling methods in a benchmarking experiment. Two were tree and forest of conditional transformation models (*R* packages *tram* and *trtf*), which detect effect modification via splitting variables (Seibold et al., 2016). The base model was a Cox Proportional Hazards (PH) regression containing the drug as the explanatory variable. The remaining methods were PH regressions regularized with an elastic net (*glmnet*) or a grouped lasso with ridge penalty (*grpreg*), which normalize the predictors during model fitting but report rescaled coefficients. The dataset for the penalized regressions included an unpenalized term for treatment, all predictors, and treatment by predictor interactions. We selected the individual predictor main terms and their treatment interactions together in the grouped lasso method.

For compatibility, we imputed the datasets for the tree and forest by a random forest method (*missForest*) but the datasets for the regularized regressions by predicted mean matching or logistic regression by chained equations (*mice*). The dataset for the grouped lasso regression was imputed once whereas we imputed the dataset for the elastic net regression five times and weighted the observations. The imputation datasets included the day of event or censoring and the Nelson-Aalen estimate of the cumulative hazard at that day. The training, test, and external validation datasets were imputed separately.

To choose the best method and parameters, we used nested 5-fold cross-validation, balanced by treatment arm. Parameter optimization was in the inner loops. For the tree, we tuned significance level for variable selection, and the minimum number of observations at a terminal node. For the forest, we tuned the number of predictors considered at split, and minimum number of observations at the terminal nodes. For the regularized regressions, we tuned the mixing parameter of lasso and ridge penalties, and the penalty parameter.

We applied the model optimized within and fitted to the training set to the outer loop, the test set, and evaluated the discrimination by cumulative time-dependent area under the curve (AUC(t)) (Bansal and Heagerty, 2018) between baseline (day 0) and days 180, 360, and 720. As a sensitivity analysis, we estimated time-dependent Brier score (Brier (t)). We chose the modeling method with the highest average AUC(t) in the test sets. We generated the final prediction model by fitting the chosen method to the whole development sample.

Any variable selected by more than one outer cross-validation folds were recorded for that modeling method, and those deemed important by more than two methods were labeled important. Only for random forest, we assessed permutation-based importance and considered the predictors with a log-likelihood importance greater than the absolute value of its minimum as important. We refrain from reporting effect estimates because the penalized nature or tree structure of the models make quantitative interpretation misleading.

2.4. Validation

With the final models, the probabilistic prognostic and treatment response predictions in the external validation sample were calculated. We estimated the 24-month AUC(t) with 95% confidence interval (CI) and plotted model and noninformative model Brier(t). We plotted, overall and stratified, 24-month receiver operating characteristic (ROC), and calibration curves. We estimated calibration-in-the-large by the intercept in a Poisson regression of actual outcome adjusting for the expected event numbers until censoring; and calibration slope by the linear predictor in a Cox PH regression of the actual outcome adjusting for the baseline hazard. Via decision curve analysis (DCA), we evaluated the net benefit of the model compared to the blanket decisions of intervention to all or no patients. We investigated the heterogeneity in response to fingolimod by treatment effect curves and measures like proportion recommended fingolimod, and variance in predicted treatment effect (Janes et al., 2014).

3. Results

There were 843 and 713 participants in the development and validation samples, respectively 425 and 358 of whom were randomized to fingolimod 0.5 mg and the rest to placebo. During follow-up, respectively 331 (39%) and 235 (33%) participants experienced a relapse, thus EPV for development was 1.9. A 24-month visit was not recorded for 112 (13%) and 148 (21%) participants in the development and validation samples. The KM curves in both samples revealed significantly higher probability of staying relapse or new T2 MRI lesion free under fingolimod compared to placebo (Fig. 2, Table S3, Figure S1). No significant difference between the arms was observed for the remaining outcomes.

At baseline in both populations (Table S4) over 70% were female and in their 30s or 40s, the median EDSS score was 2, and the median number of relapses during the 2 years prior was 2. Compared to the development, participants in the validation population were older and had a longer disease duration, were more likely to have used other DMTs, had a lesser MRI lesion load, but more comedications and concomitant diseases. On average, 0.3% of the values were missing per predictor in both datasets. Respectively in development and validation datasets, 16% and 6% of the participants had at least one missing value (Figure S2). There were no noticeable baseline characteristic differences between the treatment arms within datasets.

3.1. Development

The elastic net, with five terms, or grouped lasso, with 65 terms in their final models had the best discrimination for predicting relapse (AUC(t) 0.69) in cross-validation (Table 1). For parsimony, elastic net was chosen, the optimized parameter of which showed the use of only lasso penalization (Table S5). The final model contained only main terms: total EDSS score, total volume of Gd-enhanced T1 MRI lesions, number of relapses in the last 2 years, and number of prior MS treatments (SI4). Additionally, total volume of T2 MRI lesions, and concomitant metabolism and nutrition disorders were important predictors of relapse.

Elastic net performed similarly to grouped lasso in predicting T2 MRI lesions (AUC(t) 0.71), and immunosuppressant safety (AUC(t) 0.60), but was again chosen as final due to parsimony. In predicting 3m-CDP (AUC (t) 0.67) and the safety outcomes (AUC(t) 0.54), the transformation forest outperformed others. According to Brier(t), the chosen modeling methods would be the same, except for the safety outcome, for which elastic net (0.084) was marginally better than transformation forest (0.085) (Table S6).

Total volume and number of Gd-enhanced T1 MRI lesions, and total volume of T2 MRI lesions were the most important predictors of new T2 MRI lesions. For predicting 3m-CDP, 9-hole peg test (9HPT) and concomitant musculoskeletal and connective tissue disorders were important. For predicting safety concomitant gastrointestinal disorders, and for immunosuppressant safety, comedications of genitourinary system and sex hormones were important.

3.2. Validation

Median individual prediction for 24-month relapse risk was 0.42 (range 0.21–0.87) (Table 2). The AUC of 0.68 (95% CI: 0.63–0.72) for the relapse model was close to the cross-validation (Table 1). The Brier (t) from the model was greater than from the noninformative model but not significantly (Figure S3). The calibration plot (Fig. 3) and calibration-in-the-large (-0.17, -0.3 - -0.04) revealed risk overestimation while the calibration slope (1.06, 0.78–1.35) was acceptable (Table S7). According to the DCA, basing decisions on this model would be beneficial between the 24-month risk thresholds of 20–50% (Fig. 4). Patients, who prefer avoiding a relapse with less than 20% risk, given the risks and costs of treatment, should be treated, whereas those that require at least 50% risk should omit treatment. Treatment effect curves revealed an absence of qualitative heterogeneity (Fig. 5, Figure S4). Daily fingolimod was predicted to be more beneficial than placebo for all patients in this dataset. Median predicted individual reduction in 24-month relapse risk by fingolimod compared to placebo was 0.25 (range 0.21–0.31), with a very low variance (0.001), indicating lack of predictable treatment effect heterogeneity (Table S8).

New or enlarging T2 MRI lesion predictions had an AUC of 0.74 (95% CI 0.70–0.78) and significant improvement in Brier score. The calibration-in-the-large (-0.08, -0.18–0.01) and the calibration slope (1.07, 0.83–1.31) were acceptable. The range of risk thresholds for which the prediction model was beneficial was wide but high (40–90%, Figure S5). The variability of predicted treatment response was very low (0.001). There was a lack of treatment response heterogeneity and all patients would be recommended fingolimod. With 0.29 (range 0.12–0.32), the highest median predicted individual risk reduction by fingolimod was for this outcome.

With an AUC of 0.59 (95% CI 0.54–0.64), the CDP model performed worse than expected, and overestimated the risk: calibration-in-the-large 0.17 (0.02–0.32). The predictions were beneficial in a narrow threshold range (25–35%). The median predicted individual risk reduction in response to fingolimod was null and although only 52%

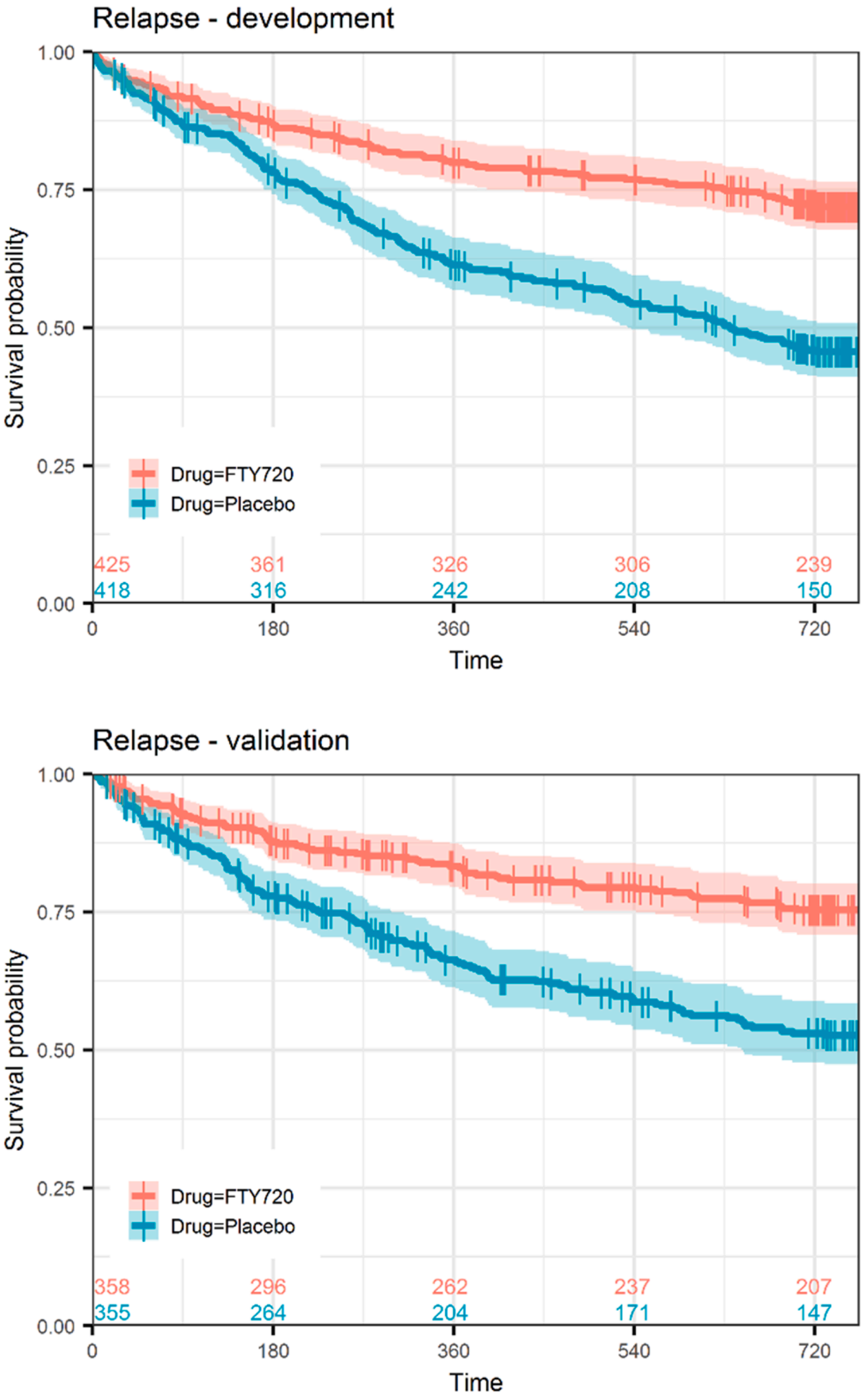


Fig. 2. Kaplan-Meier curves for relapse in development (left) and external validation (right) datasets
Survival probability is presented as a function of time in days per trial arm: active fingolimod 0.5 mg as FTY720 and control arm as Placebo. Numbers above the x-axis represent patients still under risk every 6 months.

Table 1
Time-dependent area under the curve (AUC(t)) and number of predictors.

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety
Transformation tree	0.50 (3)	0.47 (3)	0.54 (0)	0.51 (2)	0.54 (2)
Transformation forest	0.64	0.68	**0.67**	**0.54**	0.60
Elastic net	**0.69 (5)**	**0.71 (9)**	0.56 (11)	0.51 (2)	**0.60 (45)**
Grouped lasso	0.69 (65)	0.71 (19)	0.55 (35)	0.50 (17)	0.60 (81)
External Validation	0.68	0.74	0.59	0.50	0.69

First four rows: Average AUC(t) at 6, 12, and 24 months estimated via cross-validation in the model development dataset for competing methods. In parenthesis are the number of splits in interaction with treatment (transformation tree) or the terms (elastic net and grouped lasso) in the model fits. The transformation forest algorithm does not select variables, hence had all 80 predictors in interaction with treatment. The finally chosen methods are in bold. Last row: AUC(t) at 24 months estimated for the final models in the external validation dataset. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety.

Table 2
Summary of prognostic and treatment response predictions in external validation.

Outcome	Overall Event	Event in FTY720 arm	Event in Placebo arm	Treatment response
Relapse	0.42 (0.21–0.87)	0.28 (0.21–0.71)	0.53 (0.42–0.87)	0.25 (0.21–0.31)
New/enlarging lesions	0.68 (0.38–0.98)	0.47 (0.38-0.87)	0.76 (0.69–0.98)	0.29 (0.12–0.32)
Confirmed disability progression	0.22 (0.10–0.37)	0.23 (0.10–0.34)	0.22 (0.11–0.37)	0.00 (-0.18–0.18)
Safety	0.16 (0.06–0.34)	0.16 (0.06–0.33)	0.16 (0.07–0.34)	0.00 (-0.23–0.23)
Immunosuppressant safety	0.86 (0.59–1.00)	0.86 (0.59–1.00)	0.86 (0.70–0.99)	0.00 (-0.13–0.12)

First three columns: Median (range) predicted individual event probabilities at 24 months, overall and by treatment arms; FTY720: fingolimod. Last column: Median (range) of predicted individual response to fingolimod at 24 months, calculated in a counterfactual manner in all participants by predicting the risk of outcome when the treatment is fingolimod 0.5 mg and taking its difference from the predicted risk of outcome when the drug is placebo.

(48–56%) of the participants would be recommended fingolimod, overlapping CIs of the risk curves precludes a significant treatment response heterogeneity.

The safety model had an AUC of 0.50 (95% CI 0.44–0.55), indicating no discriminative power, so further results are omitted. The immunosuppressant safety was predicted with an AUC of 0.69 (0.63–0.74), although the model was significantly miscalibrated given by the calibration-in-the-large (-0.15, -0.24 - -0.07) and the calibration slope (0.66, 0.43–0.89). The model use seemed relevant for decisions with a risk threshold of 75%. The model revealed a non-significant qualitative heterogeneity where 49% (45–53%) of the participants would be recommended fingolimod and the median predicted individual risk reduction was null.

4. Discussion

We developed an easy-to-report and -implement prognostic model predicting 2-year relapse risk in RRMS patients. It is parsimonious with five main effect terms routinely collected in neurology clinics. This model had a moderate discrimination in external validation (AUC 0.68). Our result is similar to the internally validated *c*-statistics of 0.65 (Chalkou et al., 2021a), of 0.62 (Chalkou et al., 2021b), and of 0.65 (Stühler et al., 2020) from models predicting relapse in similar studies. Our model overestimated the risk and may need recalibration. Basing decisions on this model would be useful when the threshold for decision

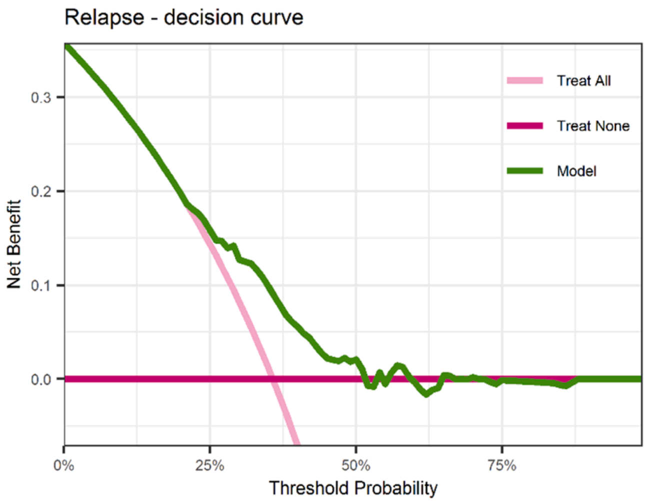


Fig. 4. Decision curve analysis in external validation
Expected net benefit, in units of proportion of true positives, under different strategies of intervention to all, none, or by the final relapse model across the entire probabilistic threshold range.

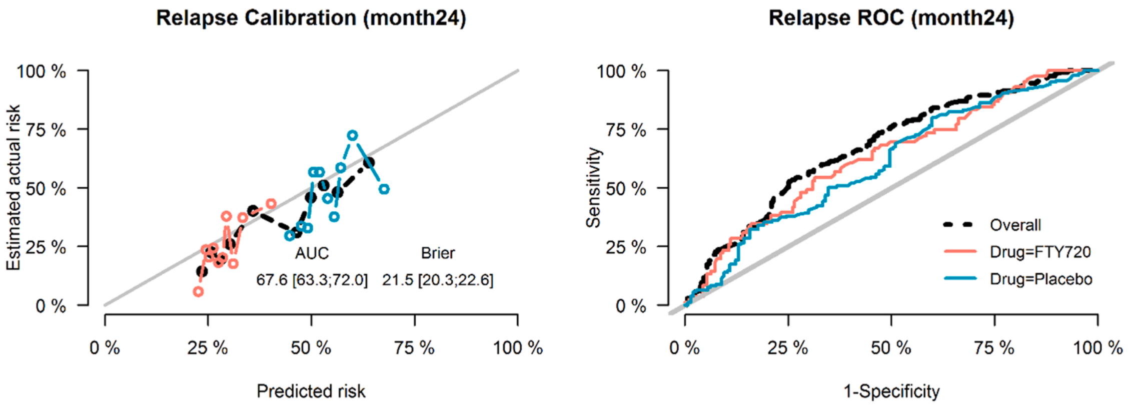


Fig. 3. Calibration and receiver operator characteristic (ROC) plots in external validation
Calibration plot binned to ten predicted risk groups (left) and ROC curve (right) overall and stratified by trial arm (active fingolimod 0.5 mg arm as FTY720 and control arm as Placebo) at month 24. Also provided are area under the curve (AUC) and Brier score as percentages.

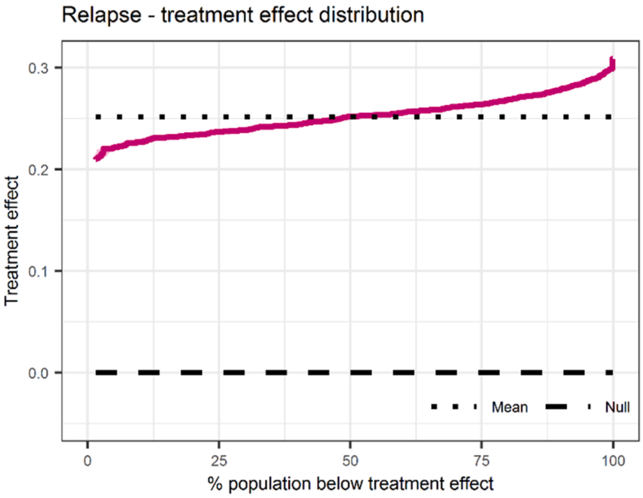


Fig. 5. Predicted treatment response in external validation
Distribution of individual treatment effect predicted counterfactually by the final model. Predicted risk under daily fingolimod 0.5 mg is considered to be the treatment from which the predicted risk under placebo is subtracted to find the treatment effect.

making is between 20% and 50% 2-year relapse risk, covering almost three-fourths of the predicted event probabilities. Previously investigated decision-relevant relapse risk thresholds are 4–25% for first line DMTs, and 19–40% for natalizumab (Chalkou et al., 2023). We expect the threshold for fingolimod to cover a similar range and our prediction model to have net benefit.

There was no predictable heterogeneity in treatment response. The individual absolute risk of 24-month relapse was predicted to be 21–31% lower with fingolimod compared to placebo for all patients in the validation. Others similarly reported lack of qualitative heterogeneity for marketed DMTs. In a meta-analysis, the main terms for the three studied DMTs and the developed prognostic score were statistically significant but their interaction was not (Chalkou et al., 2021b). Another study that included fingolimod alongside five other DMTs found less than 0.01 difference in cross-validated *c*-statistic from the model with main terms and the model with treatment interaction terms (Stühler et al., 2020).

The four selected predictors had the same direction of effect and were selected by another relapse risk model (Chalkou et al., 2021b). Our results also confirm the findings from the FREEDOMS trials' subgroup analysis that patients with higher baseline disease activity or disability had higher relapse rates (Derfuss et al., 2016). These predictors represent disability (total EDSS score) and symptoms (number of relapses in the last 2 years). Higher total volume of Gd-enhanced T1 MRI lesions, which detect new inflammatory activity, also decreased the time-to-relapse. Exposure to more DMTs pre-baseline increased the relapse risk. Other important predictors were total volume of T2 MRI lesions and concomitant metabolism and nutrition disorders. Age or sex were not found to be influential independent predictors by our multivariable methods.

The outcome predicted with best performance, new or enlarging T2 MRI lesions, had an AUC of 0.74. There was no predictable heterogeneity also for this outcome in response to fingolimod. Its most important positively correlated predictors were other lesion markers like volume and number in T1 MRI and volume in T2 MRI.

The 3m- CDP was predicted more successfully by the transformation forest, indicating higher-order interactions to be explanatory for disability. Its AUC in external validation (0.59) was comparable to internal validation from similar studies: worse than 0.65 for disability progression (Pellegrini et al., 2020b) and 0.69 for CDP (De Brouwer et al., 2021; De Brouwer et al., 2022), but better than 0.56 for another CDP model (Stühler et al., 2020). The null median predicted treatment response was in agreement with the source trial results, FREEDOMS II (Calabresi et al., 2014). Hand dexterity (9HPT), and concomitant musculoskeletal and connective tissue disorders were the most important predictors of 3m-CDP. Also important in the forest were EDSS total and cerebral function scores. Lesion number or volume in T1 or T2 MRI were not important in any of the methods, hinting at the "clinico-radiological paradox", that disease activity and disability progression may have different mechanisms (Barkhof, 2002).

A novelty of our study was predicting safety-related outcomes. The overall safety could not be modeled with sufficient discrimination, pointing to the tension in their modeling: the difficulty in capturing varying underlying mechanisms when the grouping is coarse versus the rarity of SAEs. The outcome of infections and infestations and neoplasms could be better predicted (AUC 0.69). For this outcome, baseline lymphocyte count was one of twenty-nine important predictors selected by the regression models.

In terms of predicting the treatment response heterogeneity, randomization is the unbiased way and RCT data is completer and more standardized. Yet, there are barriers to this prediction, like the impossibility of observation, lack of known strong effect modifiers, and that RCTs are underpowered to detect multiple weak interactions (Rekkas et al., 2020). There is a strong deterministic assumption underlying the discourse on treatment response heterogeneity: the variability in the outcomes observed in the active arm in an RCT is a quality of the patient and cannot vary within the patient in a random manner (Senn, 2018). Our study based on (para)clinical parameters shows that this assumption is questionable for the relapse outcome in response to fingolimod.

In terms of prognosis, despite the high quality of the dataset on which they are based on, our models are limited and are very early in their developmental stage. They can only be used to predict risk under no treatment, i.e. placebo, or fingolimod. Even before predicting the individual prognosis of relapse under no treatment, which can support the decision on how strong of a DMT to use, our models are not yet ready for implementation. RCT participants are expected to have higher disease activity and less comorbidities compared to the overall RRMS patients (Trojano et al., 2017), hence the prediction models developed in this study require validation in, and maybe calibration to, routine datasets before any clinical implementation.

5. Conclusions

The developed and externally validated prognostic model predicting relapse under no treatment or fingolimod should be further externally validated, recalibrated, and tested in impact studies. Only after these stages are successful, it can be implemented as a tool and used for decision support in clinical practice. We found no predictable heterogeneity of disease activity in response to fingolimod. Fingolimod would be a good treatment option to RRMS patients if 21% absolute relapse risk reduction compared to placebo outweighs its possible safety risks. Further research is warranted to investigate our exploratory findings, namely the prognostic value of concomitant diseases in predicting clinical outcomes in RRMS.

Data and code availability

The data that support the findings of this study are available from the data owner, Novartis, upon reasonable request via appropriate platforms. The code is publicly available at github.com/irmakon/FTY720_TreatmentResponse.

Funding

This work was supported by the DIFUTURE project funded by the German Federal Ministry of Education and Research under 01ZZ1804B and 01ZZ1804C. The funding source did not have any role in study

design; in the collection, analysis and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

CRediT authorship contribution statement

Begüm Irmak Ön: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Joachim Havla:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing. **Ulrich Mansmann:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

BIÖ and UM have nothing to declare. JH reports a grant for OCT research from the Friedrich-Baur-Stiftung, Horizon, Sanofi and Merck, personal fees and nonfinancial support from Alexion, Amgen, Bayer, Biogen, BMS, Merck, Novartis and Roche, and nonfinancial support of the Sumaira-Foundation and Guthy-Jackson Charitable Foundation, all outside the submitted work.

Acknowledgments

This work was supported by the DIFUTURE project funded by the German Federal Ministry of Education and Research under 01ZZ1804B and 01ZZ1804C. Novartis, the sponsor of the FREEDOMS trials shared the anonymous study data, and Clinical Study Data Request provided the analysis platform. Prof. Dr. Ulrike Held and Dr. Kelly Reeve contributed to the conceptualization of this study. Josef Herker contributed to the curation of the analysis dataset.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.msard.2024.106247.

References

Baecher-Allan, C., Kaskow, B.J., Weiner, H.L., 2018. Multiple sclerosis: mechanisms and immunotherapy. Neuron 97 (4), 742–768.

Bansal, A., Heagerty, P.J., 2018. A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. Med. Decis. Making 38 (8), 904–916.

Barkhof, F., 2002. The clinico-radiological paradox in multiple sclerosis revisited. Curr. Opin. Neurol. 15 (3), 239–245.

Brown, F.S., Glasmacher, S.A., Kearns, P.K.A., MacDougall, N., Hunt, D., Connick, P., Chandran, S., 2020. Systematic review of prediction models in relapsing remitting multiple sclerosis. PLoS One 15 (5), e0233575.

Calabresi, P.A., Radue, E.-W., Goodin, D., Jeffery, D., Rammohan, K.W., Reder, A.T., Vollmer, T., Agius, M.A., Kappos, L., Stites, T., Li, B., Cappiello, L., von Rosenstiel, P., Lublin, F.D., 2014. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. Lancet Neurol 13 (6), 545–556.

Chalkou, K., Steyerberg, E., Bossuyt, P., Subramaniam, S., Benkert, P., Kuhle, J., Disanto, G., Kappos, L., Zecca, C., Egger, M., Salanti, G., 2021a. Development, validation and clinical usefulness of a prognostic model for relapse in relapsing-remitting multiple sclerosis. Diagn. Progn. Res. 5 (1), 17.

Chalkou, K., Steyerberg, E., Egger, M., Manca, A., Pellegrini, F., Salanti, G., 2021b. A two-stage prediction model for heterogeneous effects of treatments. Stat. Med. 40 (20), 4362–4375.

Chalkou, K., Vickers, A.J., Pellegrini, F., Manca, A., Salanti, G., 2023. Decision curve analysis for personalized treatment choice between multiple options. Med. Decis. Making 43 (3), 337–349.

De Brouwer, E., Becker, T., Moreau, Y., Havrdova, E.K., Trojano, M., Eichau, S., Ozakbas, S., Onofrj, M., Grammond, P., Kuhle, J., Kappos, L., Sola, P., Cartechini, E., Lechner-Scott, J., Alroughani, R., Gerlach, O., Kalincik, T., Granella, F., Grand'Maison, F., Bergamaschi, R., José Sá, M., Van Wijmeersch, B., Soysal, A., Sanchez-Menoyo, J.L., Solaro, C., Boz, C., Iuliano, G., Buzzard, K., Aguera-Morales, E., Terzi, M., Trivio, T.C., Spitaleri, D., Van Pesch, V., Shaygannejad, V., Moore, F., Oreja-Guevara, C., Maimone, D., Gouider, R., Csepany, T., Ramo-Tello, C., Peeters, L., 2021. Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. Comput. Methods Programs Biomed. 208, 106180.

De Brouwer, E., Becker, T., Moreau, Y., Havrdova, E.K., Trojano, M., Eichau, S., Ozakbas, S., Onofrj, M., Grammond, P., Kuhle, J., Kappos, L., Sola, P., Cartechini, E., Lechner-Scott, J., Alroughani, R., Gerlach, O., Kalincik, T., Granella, F., Grand`Maison, F., Bergamaschi, R., Sá, M.J., Van Wijmeersch, B., Soysal, A., Sanchez-Menoyo, J.L., Solaro, C., Boz, C., Iuliano, G., Buzzard, K., Aguera-Morales, E., Terzi, M., Trivio, T.C., Spitaleri, D., Van Pesch, V., Shaygannejad, V., Moore, F., Oreja-Guevara, C., Maimone, D., Gouider, R., Csepany, T., Ramo-Tello, C., Peeters, L., 2022. 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 213, 106479.

Derfuss, T., Ontaneda, D., Nicholas, J., Meng, X., Hawker, K., 2016. Relapse rates in patients with multiple sclerosis treated with fingolimod: subgroup analyses of pooled data from three phase 3 trials. Mult. Scler. Relat. Disord. 8, 124–130.

Devonshire, V., Havrdova, E., Radue, E.W., O'Connor, P., Zhang-Auberson, L., Agoropoulou, C., Häring, D.A., Francis, G., Kappos, L., 2012. Relapse and disability outcomes in patients with multiple sclerosis treated with fingolimod: subgroup analyses of the double-blind, randomised, placebo-controlled FREEDOMS study. Lancet Neurol 11 (5), 420–428.

Havas, J., Leray, E., Rollot, F., Casey, R., Michel, L., Lejeune, F., Wiertlewski, S., Laplaud, D., Foucher, Y., 2020. Predictive medicine in multiple sclerosis: a systematic review. Mult. Scler. Relat. Disord. 40, 101928.

Holstiege, J., Akmatov, M.K., Klimke, K., Dammertz, L., Kohring, C., Marx, C., Frahm, N., Peters, M., Ellenberger, D., Zettl, U.K., Rommer, P.S., Stahmann, A., Bätzing, J., 2022. Trends in administrative prevalence of multiple sclerosis and utilization patterns of disease modifying drugs in Germany. Mult. Scler. Relat. Disord. 59, 103534.

Janes, H., Brown, M.D., Huang, Y., Pepe, M.S., 2014. An approach to evaluating and comparing biomarkers for patient treatment selection. Int. J. Biostat. 10 (1), 99–121.

Kappos, L., Radue, E.-W., O'Connor, P., Polman, C., Hohlfeld, R., Calabresi, P., Selmaj, K., Agoropoulou, C., Leyk, M., Zhang-Auberson, L., Burtin, P., 2010. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. N. Engl. J. Med. 362 (5), 387–401.

Kent, D.M., Steyerberg, E., Klaveren, D.v., 2018. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ 363.

Moons, K.G.M., Wolff, R.F., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J., Mallett, S., 2019. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann. Intern. Med. 170 (1), W1.

Müller, S., Heidler, T., Fuchs, A., Pfaff, A., Ernst, K., Ladinek, G., Wilke, T., 2020. Real-world treatment of patients with multiple sclerosis per MS subtype and associated healthcare resource use: an analysis based on 13,333 patients in Germany. Neurol. Ther. 9 (1), 67–83.

Ohlmeier, C., Gothe, H., Haas, J., Osowski, U., Weinhold, C., Blauwitz, S., Schmedt, N., Galetzka, W., Berkemeier, F., Tackenberg, B., Stangel, M., 2020. Epidemiology, characteristics and treatment of patients with relapsing remitting multiple sclerosis and incidence of high disease activity: Real world evidence based on German claims data. PLoS One 15 (5), e0231846.

Pellegrini, F., Copetti, M., Bovis, F., Cheng, D., Hyde, R., de Moor, C., Kieseier, B.C., Sormani, M.P., 2020a. A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. Mult. Scler. 26 (9), 1064–1073.

Pellegrini, F., Copetti, M., Sormani, M.P., Bovis, F., de Moor, C., Debray, T.P.A., Kieseier, B.C., 2020b. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. Mult. Scler. 26 (14), 1828–1836.

Reeve, K., On, B.I., Havla, J., Burns, J., Gosteli-Peter, M.A., Alabsawi, A., Alayash, Z., Götschi, A., Seibold, H., Mansmann, U., Held, U., 2023. Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. https://doi.org/10.1002/14651858.CD013606.pub2.

Rekkas, A., Paulus, J.K., Raman, G., Wong, J.B., Steyerberg, E.W., Rijnbeek, P.R., Kent, D.M., van Klaveren, D., 2020. Predictive approaches to heterogeneous treatment effects: a scoping review. BMC Med. Res. Methodol. 20, 264.

Seibold, H., Zeileis, A., Hothorn, T., 2016. Model-based recursive partitioning for subgroup analyses. Int. J. Biostat. 12 (1), 45–63.

Senn, S., 2018. Statistical pitfalls of personalized medicine. Nature 563 (7733), 619–621.

Steyerberg, E.W., Moons, K.G.M., Windt, D.A.v.d., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G., Group, f.t.P., 2013. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 10 (2), e1001381.

Stühler, E., Braune, S., Lionetto, F., Heer, Y., Jules, E., Westermann, C., Bergmann, A., van Hövell, P., NeuroTransData Study, G., 2020. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. BMC Med. Res. Methodol. 20 (1), 24.

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

Wang, R., Lagakos, S.W., Ware, J.H., Hunter, D.J., Drazen, J.M., 2007. Statistics in medicine — reporting of subgroup analyses in clinical trials. N. Engl. J. Med. 357 (21), 2189–2194.

Wiendl, H., Gold, R., Berger, T., Derfuss, T., Linker, R., Mäurer, M., Aktas, O., Baum, K., Berghoff, M., Bittner, S., Chan, A., Czaplinski, A., Deisenhammer, F., Di Pauli, F., Du Pasquier, R., Enzinger, C., Fertl, E., Gass, A., Gehring, K., Gobbi, C., Goebels, N., Guger, M., Haghikia, A., Hartung, H.-P., Heidenreich, F., Hoffmann, O., Kallmann, B., Kleinschnitz, C., Klotz, L., Leussink, V.I., Leutmezer, F., Limmroth, V., Lünemann, J.D., Lutterotti, A., Meuth, S.G., Meyding-Lamadé, U., Platten, M., Rieckmann, P., Schmidt, S., Tumani, H., Weber, F., Weber, M.S., Zettl, U.K., Ziemssen, T., Zipp, F., 2021. Multiple sclerosis therapy consensus group (MSTCG): position statement on disease-modifying therapies for multiple sclerosis (white paper). Ther. Adv. Neurol. Disord. 14.

Winkelmann, A., Loebermann, M., Reisinger, E.C., Hartung, H.-P., Zettl, U.K., 2016. Disease-modifying therapies and infectious risks in multiple sclerosis. Nat. Rev. Neurol. 12 (4), 217–233.

Ziemssen, T., Akgün, K., Brück, W., 2019. Molecular biomarkers in multiple sclerosis. J. Neuroinflammation 16 (1), 272.