Aus dem
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE)
Ludwig-Maximilians-Universität München

Prediction of prognosis and response to fingolimod in people with relapsing-remitting multiple sclerosis

Dissertation
zum Erwerb des Doctor of Philosophy (Ph.D.)
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von
Begüm Irmak Ön

aus
Üsküdar, İstanbul, Türkei

Jahr
2023

Mit Genehmigung der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

Erstes Gutachten:	Prof. Dr. Ulrich Mansmann
Zweites Gutachten:	Prof. Dr. Martin Kerschensteiner
Drittes Gutachten:	Priv. Doz. Dr. Markus Krumbholz
Viertes Gutachten:	Prof. Dr. Andrea Szelényi

Dekan: Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung: 13.11.2023

Table of content

Abstract

Multiple sclerosis (MS) is a multifactorial neurological condition that is progressive and disabling. During the last decades, over 12 treatments with varying mechanisms were marketed for the relapsing-remitting phenotype, the most common subtype at MS diagnosis. The unpredictability of the individual disease courses and response to treatments is a challenge routinely faced in clinical practice. Individualized prognostic and treatment response prediction by multivariable models can support medical decisions. Well-conducted prediction studies in MS indicate poor to moderate predictability of efficacy outcomes.

To complement the prognostic literature in MS, this study aimed to predict response to fingolimod and identify important prognostic predictors. Time-to relapse and other efficacy and safety endpoints were predicted by repurposing placebo and fingolimod 0.5 mg arms from two randomized controlled trials. Models based on Cox proportional hazards were developed with data from the FREEDOMS trial (n=843) by allowing transformation tree, transformation forest, elastic net, and grouped lasso methods to compete in a nested cross-validation. In addition to the treatment arm, 80 baseline predictors and treatment by predictor interactions were considered as candidate variables in the models. Reproducibility of the models with the highest cross-validated area under the receiver operating curve (AUC) were evaluated by external validation with the data from the FREEDOMS II trial (n=713).

The final model predicting relapse risk was an elastic net regression with main terms for treatment and four predictors. In the external validation, it had a moderate two-year AUC of 0.68 (95% confidence interval: 0.63-0.72), but the predictions were overestimating the actual risk. There was almost no heterogeneity in the predicted treatment response (variability 0.001) and all participants were predicted to have 0.21 to 0.31 absolute relapse risk reduction with fingolimod compared to placebo. The final model predicting new or enlarging T2 magnetic resonance imaging (MRI) lesions had an AUC of 0.74 (0.70-0.78), moderate calibration, but lack of variability in the predicted treatment response. The model predicting confirmed disability progression had an AUC of 0.59 (0.54-0.64) and non-significant heterogeneity in the predicted treatment response. The safety outcome of serious adverse events or trial discontinuation was not predictable with sufficient discrimination. The model predicting infections or neoplasms had an AUC of 0.69 (0.63-0.74), but poor calibration and non-significant heterogeneity in the predicted treatment response. Many important predictors of the efficacy outcomes were related to (para)clinical disease activity or disability. Unexpected influential predictors included concomitant metabolism and nutrition disorders for relapse, musculoskeletal and connective tissue disorders for confirmed disability progression, and gastrointestinal disorders for safety.

The two-year predictability of relapse and new or enlarging lesions in T2 MRI were moderate to good and the predicted change in their risk as response to fingolimod was a decrease for all patients, lacking heterogeneity. Following further satisfactory external validations, models for these disease activity outcomes can be used for prognostic prediction in clinical care. The predictability of disability and safety outcomes were poor and it remains unclear whether the change in their risk as response to fingolimod is heterogeneous.

List of figures

List of tables

List of abbreviations

Abbreviation	Term
9HPT	Nine-hole peg test
ARR	Annualized relapse rate
ATC	Anatomical therapeutic chemical
AUC	Area under the (receiver operating) curve
CDP	Confirmed disability progression
CI	Confidence interval
CIS	Clinically isolated syndrome
CNS	Central nervous system
CSF	Cerebrospinal fluid
DMT	Disease-modifying therapy
EDSS	Expanded disability status scale
EMA	European Medicines Agency
EPV	Events per variable
FDA	Food and Drug Administration
Gd	Gadolinium
HR	Hazard ratio
ITT	Intention-to-treat
MedDRA	Medical dictionary for regulatory activities
MRI	Magnetic resonance imaging
MS	Multiple sclerosis
MSFC	Multiple sclerosis functional composite
PASAT	Paced auditory serial addition test
PH	Proportional hazards
PPMS	Primary progressive multiple sclerosis
PROBAST	Prediction model risk of bias assessment tool
QoL	Quality of life
RCT	Randomized controlled trial
RRMS	Relapsing-remitting multiple sclerosis
S1P	Sphingosine-1-phosphate
SAE	Serious adverse event
SF-36	36-item short-form health survey
SOC	System organ class
SPMS	Secondary progressive multiple sclerosis
SUCRA	Surface under the cumulative ranking
T25FW	Timed 25-foot walk
VFT	Visual function test

1. Introduction

This monographic dissertation, reported according to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis statement[1], starts with setting the medical (Section 1.1) and methodological (Section 1.2) context of the research question in this introductory Chapter. The current knowledge on the epidemiology and etiology of the health condition of interest, multiple sclerosis, is summarized in Section 1.1.1. This is followed by the introduction of pharmaceutical treatment options and relevant clinical endpoints in Section 1.1.2, and going deeper into the treatment of interest, fingolimod, in Section 1.1.3. A similar narrowing approach is followed to introduce prognostic and treatment response prediction starting with its definition and importance in Section 1.2.1. The state-of-the-art prediction-relevant methods of development and validation are summarized in 1.2.2 and the current state of the scientific literature on prognostic and treatment response prediction in multiple sclerosis is summarized in Section 1.2.3. The bottom-line of this Chapter and the gap in the literature, which this thesis aims to fill, are summarized in Section 1.3.

1.1 Multiple sclerosis

1.1.1 Disease and epidemiology

Multiple sclerosis (MS) is a chronic inflammatory disease of the central nervous system (CNS), the main sign of which is demyelination and axonal degeneration in the brain and the spinal cord.[2] The disease typically starts in young adulthood, between 20 to 40 years of age, and is potentially not only disabling but also progressive.[3-5] Like other autoimmune diseases that are considered to be partially linked to immunological mechanisms mediated by the X-chromosome[6], MS is two to three times more common in women and its symptoms become milder during pregnancy.[3,7] MS is the most prevalent progressively debilitating neurological condition in young adults.[8] The worldwide age-standardized prevalence of MS was estimated to be 30.1 cases (95% uncertainty interval 27.5-33.0) per 100 000 in 2016. The prevalence is highest in developed regions with 164.4 cases (153.2-177.1) in North America and 127.0 (115.4-139.6) cases in Western Europe.[9] Also, the disease prevalence has been increasing in many regions since the 1990, at least partly due to earlier and better diagnosis, and longer survival.[10] The disparity in prevalence between regions have been attributed to many factors including the latitude and sunlight exposure, other undetermined environmental exposures in the developed regions, and underdiagnosis in developing regions.[9]

The inflammatory mechanisms involved in the pathogenesis of MS are neither singular nor simple to disentangle. This is evident in the various mechanisms of action of the approved drugs for this indication. Different drugs interfere with the pathways related to effector T cells, regulatory T cells, B cells, or immune cell migration.[11] The activation of both the innate and adaptive immune systems, linked to not only hereditary but also environmental factors, have been associated with MS. Increased immune cell population in the CNS is followed by an attack to the myelin producing oligodendrocytes, which then leads to the destruction of the myelin sheath of the neuronal axons and appearance of plaques visible by neuroimaging. Additional injury to the axons and the neuronal body can occur and may be measured in the white as well as the gray matter. In the progressive forms of MS, immunity-independent mechanisms like oxidative stress are also thought to play a role.[7]

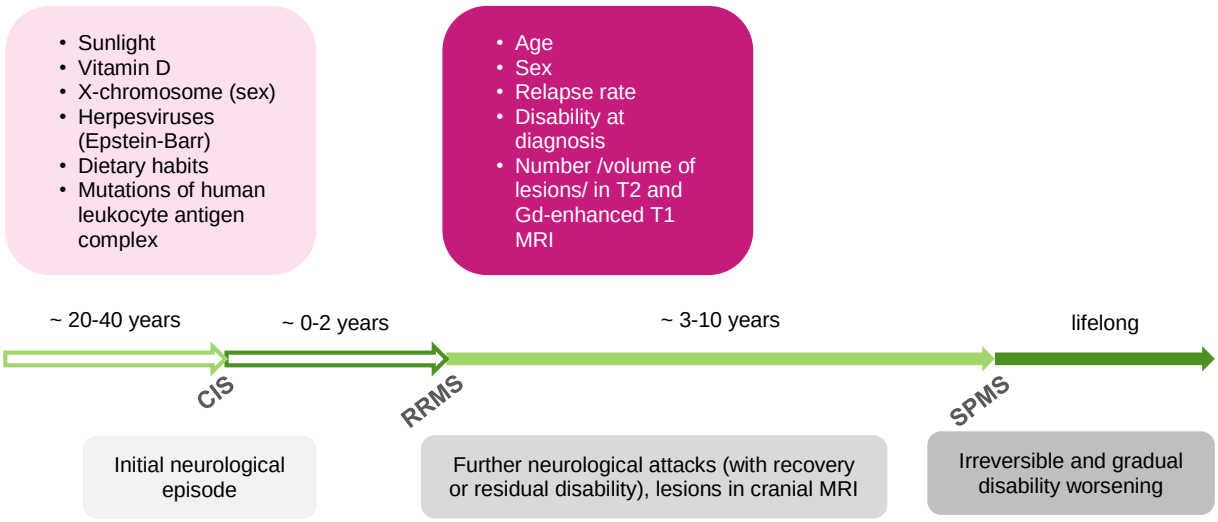


Figure 1 *Timeline and risk factors of multiple sclerosis*
A simplified disease history of a typical relapsing-remitting multiple sclerosis (RRMS) patient, who is initially diagnosed with clinically-isolated syndrome (CIS) and who progresses to secondary progressive phenotype (SPMS). Predictors of MS diagnosis (left) and prognosis (right) according to the current literature are represented above the timeline. MRI: magnetic resonance imaging.

Being an autoimmune disease, MS is thought to manifest in genetically susceptible people who are exposed to triggers.[12] The genetic loci of the human leukocyte antigen complex have been identified as the most influential on the risk of MS development. Other identified susceptibility loci, mutations of which probably act through epigenetic mechanisms, have been also related to immune cell function.[7] Immune activation via exposure to herpesviruses (particularly Epstein-Barr virus), arguably at a certain developmental stage, and dietary habits affecting the microbiome and metabolism have been linked to the probability of developing MS. In contrast, vitamin D and exposure to sunlight required for its conversion to an active form, have been shown to exert a protective effect against developing MS. Taking its multifactorial etiology into account[11], the current state of the research indicates that, like many complex diseases, MS is the result of an interaction between genetic predisposition and environment.[13] However, it is unclear whether the genetic component can explain the heterogeneity in disease trajectory or its mildness whereas environmental factors seem to influence the disease also post-diagnosis.[11]

Although MS is thought to have a considerable initial subclinical phase[11], its diagnosis necessitates clinical ques.[14-18] The disease clinically manifests by acute neurological attacks affecting one or more functional systems and gradual development or worsening of disability. The development of neuroimaging techniques has led to the identification of disease activity by observing not only clinical symptoms but also lesions (scleroses) within the CNS.[17,19] MS is categorized into subtypes based on observed signs, symptoms, and the pace of disability progression. An initial neurological attack without any obvious reason and without (para)clinical evidence of dissemination in time and space (multiplicity) to confirm an MS diagnosis is called clinically isolated syndrome (CIS), a subcategory likely to develop into MS.[18] The majority of people affected by MS initially present with a relapsing-remitting subtype (RRMS, **Figure 1**), characterized by intermittent neurological attacks and evidence of demyelination, which may recover or leave residual neurological deficits.[3] After varying number of years from RRMS onset or sometimes right from the start, the disease settles into a relatively stable lifelong course with gradual accumulation of disability, at which point it is called secondary progressive MS (SPMS) or primary progressive MS (PPMS), respectively. The disease course, its severity, and its (para)clinical

manifestations are highly heterogeneous[20,21] and largely remain unpredictable[22,23] for people with all MS subtypes, but especially for those with RRMS.

Although there are established diagnostic paraclinical and biological markers in MS[18], like the presence of gadolinium-enhanced (Gd-enhanced) T1 lesions in cranial magnetic resonance imaging (MRI) scans or absence of anti-aquaporin-4 antibodies, widely accepted prognostic and predictive markers are lacking.[21,24] Hence, risk factors for the highly active (a.k.a. aggressive, malignant) disease phenotype remain as potential candidates when the aim is prognostic or treatment response prediction. Although a consensus definition is lacking[25], the operationalization of highly active disease is usually based on severity and frequency of clinical or radiological disease activity, and speed of disability worsening.[4,26,27]

1.1.2 Treatment landscape

Even though MS is still an incurable disease, from 1995 to 2022, 14 immunotherapies have been granted marketing authorization by the European Medicines Agency (EMA).[4,28] Most of these so-called disease-modifying therapies (DMT) belong to the Anatomical Therapeutic Chemical (ATC) classification subgroup of immunosuppresants and are primarily indicated for the RRMS subtype. Initially, injectable therapies were approved. Interferon beta preparations were marketed in the 1990s, followed by glatiramer acetate and a monoclonal antibody, natalizumab, in 2000s. Fingolimod, a sphingosine-1-phosphate (S1P) receptor modulator, was the first oral treatment to be marketed for the treatment of RRMS, followed by other oral therapies such as dimethyl fumarate and teriflunomide. Starting in 2010s, more S1P receptor modulators (ozanimod, ponesimod), monoclonal antibodies (alemtuzumab, ocrelizumab, ofatumumab), and other immune-modifying therapies (cladribine) have been approved for use in the treatment of MS. This section is concentrated on the 12 DMTs approved for the RRMS phenotype because it is not only the subtype with the greatest number of treatment options but also the focus of this thesis.

In phase III clinical trials, the efficacies of the DMTs have been demonstrated with endpoints based primarily on relapses, usually operationalized as annualized relapse rate (ARR) within a time span of 1 to 3 years.[29,30] Also common are operationalizations of relapse by proportion of relapse-free participants or time-to-first relapse. Although it varies in the literature, the definition of a relapse mostly comprises of a specific timeframe (e.g. neurological symptoms lasting more than 24 hours), independence from previous relapses (e.g. 30 days' gap) or other non-MS related possible triggers (e.g. without fever or sign of infection), and relation to neurological findings or disability score changes.[31] A secondary efficacy outcome common in interventional trials of DMTs is disability worsening, which is usually defined by relapse-independent measurement of Expanded Disability Status Scale (EDSS)[32,33] that needs to be confirmed with a repeated measurement at 3 or 6 months.[31] Additionally, MRI-based outcomes of new or enlarging T2 hyperintense or Gd-enhanced T1 lesions are common secondary endpoints in pivotal trials.[33,34] However, these can be primary endpoints in phase II trials because MRI-based event rate is higher than the rate of clinical endpoints and its surrogacy to relapse rate has been demonstrated at the population level for many drugs.[35,36] Outcomes formed from different components, such as no evidence of disease activity and its variants, have also been used in efficacy studies.[31,33,37] When asked to order the importance of different outcomes, people with MS prioritize disability progression, relapses, and serious adverse events (SAE) over others. MRI changes, which are common surrogates in clinical trials, rank as the least important outcome for MS patients.[38]

Of the DMTs approved for the RRMS indication, interferon beta and glatiramer acetate, a.k.a first-line therapies, have moderate efficacy and are relatively safe. The remaining therapies (natalizumab, fingolimod, alemtuzumab, teriflunomide, dimethyl fumarate, cladribine, ocrelizumab, ozanimod, ofatumumab, ponesimod), a.k.a. second-line, are more potent leading to a higher chance of not only

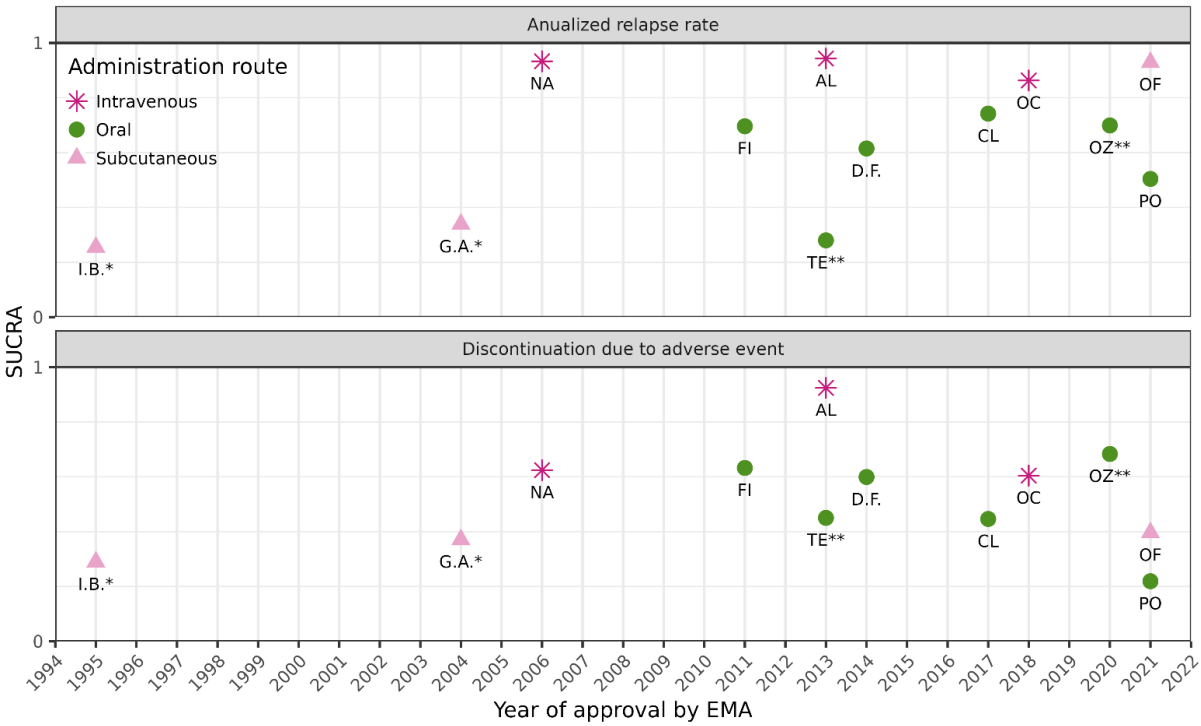


Figure 2 *Overview of multiple sclerosis treatments*

Efficacy and safety overview of disease-modifying therapies approved by the European Medicines Agency (EMA)[28] for people with relapsing-remitting multiple sclerosis. The data is based on surface under the cumulative ranking (SUCRA) probabilities in a network meta-analysis of randomized controlled trials.[39] Higher SUCRA values indicate higher effect. The SUCRA is provided for illustrative purposes only and its limitations, such as not accounting for magnitude or uncertainty of the effect, should be borne in mind. I.B.: interferon betas, G.A.: glatiramer acetate, NA: natalizumab, FI: fingolimod, AL: alemtuzumab, TE: teriflunomide, D.F.: dimethyl fumarate, CL: cladribine, OC: ocrelizumab, OZ: ozanimod, OF: ofatumumab, PO: ponesimod. *For simplicity, SUCRAs of dosages or formulations were averaged: five different formulations of Interferon-beta, subcutaneous or intramuscular, and two different approved dosages of glatiramer acetate. **Only the approved adult maintenance dosages are reported: 14 mg for teriflunomide, and 0.92 mg for ozanimod.

success in disease control but also potentially serious adverse reactions. However, this dichotomization into first or second-line is a simplification and the list of therapies exhibit a range of efficacy and safety profiles (**Figure 2**). A different framework in which the therapies are considered within three efficacy categories was incorporated into the clinical guideline by the German Neurological Society. In this framework, fingolimod is considered to be in the middle category.[27]

The current treatment guidelines and position statements recommend early initiation of DMTs in patients diagnosed with RRMS[4], especially if disease activity has been observed recently or during the last two years, to be more specific.[40] Yet, which treatment to choose in treatment-naïve patients is less clear. In case of favorable disease characteristics, not to initiate a treatment but monitoring carefully is also an option which may be preferred by the patients.[4,27] The current treatment paradigm tends to focus on two strategies. The first is the conservative step-wise escalation approach, which has been the mainstream until recently but currently is recommended for patients with low to moderate disease activity. In this strategy, the patient is initially treated with a moderate efficacy treatment until a persisting or recurring exacerbation of the disease activity is observed, after which the patient is switched to a stronger but riskier treatment.[27,41] The second strategy of induction, sometimes referred to as “hard and early” approach, is typically reserved for otherwise healthy patients with highly active disease, the operationalization of which is not universal.[27] In this strategy, the patient is treated aggressively with a

high efficacy treatment starting from diagnosis[40,41], which may lead to a better control of disease symptoms as well as serious adverse reactions. In case of a perceived lack of efficacy while on a second-line DMT, the recommendation is to switch to another second-line DMT.[4,41]

Choosing a treatment strategy boils down to prioritizing safety by escalation versus efficacy by induction.[42] Some proponents of the latter approach also advocate for de-escalation either to a milder treatment or discontinuing altogether when the disease is under control and presents a stable course for a period of time[24], especially in the elderly for whom safety concerns of the DMTs are paramount and their effectiveness questioned.[43] The German Neurological Society guideline also considers drug discontinuation when the disease is under control with moderate efficacy treatment.[27] To prevent irreversible disability while minimizing undesired effects, using the "right treatment at the right time" is the aspired objective in MS.[44]

There are some responder and some non-responder RRMS patients to all available therapies.[11] Identifying them, preferably prior to treatment, is the challenge that is yet to be overcome. Conventional subgroup analyses of clinical trials form the first step to individualizing therapies by finding effect heterogeneity in patients so that both patients and physicians are better informed during treatment decisions.[45] With the intention to find groups of patients that would benefit from a DMT irrespective of the specific treatment, subgroup analyses from clinical trials of different DMTs were meta-analyzed.[46] The effect of treatment within subgroups relative to its overall effect were statistically combined. The included subgroup analyses originated from six blinded placebo-controlled randomized controlled trials (RCT) of the drugs natalizumab, fingolimod, dimethyl fumarate, teriflunomide, and glatiramer acetate. The relative treatment effect on relapse rate was significantly greater in younger (less than 40 years of age in all except less than 38 in one study) participants compared to older participants. Baseline Gd-enhanced MRI activity, and EDSS lower than 3.5 were also found to be significantly interacting with the treatments' effect on relapse rate. In the meta-analysis for the outcome of disability progression, the only factor significantly interacting with the treatment effect was found to be age.

The safety of DMTs may also be heterogeneous. Different therapies may have adverse effects in some of the patients but not others. Owing to their main therapeutic action of immune suppression, all DMTs induce a risk of leukopenia/lymphopenia and many of them also introduce an increase in risk of infections and neoplasms to varying degrees.[4,47,48] The degree of the risk and its particularities (like category of infection) depend on the mode of action of the individual therapies.[47]

Despite the availability of many DMT options[49], small number of head-to-head comparisons leads to difficulty in treatment decision making. Recent network meta-analyses based on systematic reviews of RCTs[30,39,50] perpetuate the existing knowledge that the intravenous treatments (alemtuzumab, natalizumab, and ocrelizumab) have the highest efficacy, but fail to agree about the acceptability of DMTs. The multiplicity and recency of treatment options with different modes of action and safety-efficacy profiles[49,51] have been the motivation for researchers in the MS field to investigate ways for supporting treatment decisions by scoring or predicting treatment response, in particular treatment efficacy and effectiveness.

1.1.3 Fingolimod

Fingolimod is the first oral DMT approved in 2010 by the Food and Drug Administration (FDA) in the United States and in 2011 by the EMA in Europe. It is considered to be a continuously-used treatment option with high efficacy[4] for the treatment of patients with RRMS. In the US, fingolimod is used as first-line treatment[52], whereas in the EU, it is indicated as second-line after treatment failure with other therapies or first-line in patients with a rapidly evolving disease course.[53] As an S1P receptor modulator,

Study	Eligibility criteria at baseline	Arms	Patients (n, median)	Results
FREEDOMS[54]	• RRMS (McDonald 2005 criteria) • aged 18-55y • EDSS <6.0 • 1 relapse in 1y or 2 relapses in 2y pre-randomization • 3m gap after interferon beta or glatiramer acetate • 6m gap after natalizumab	1:1:1 daily placebo (comparator) daily fingolimod 0.5 mg daily fingolimod 1.25 mg	1272 participants (ITT) 1033 completed the trial age 37y disease duration 6.7y EDSS score 2.0	At 2y ARR reduction fingolimod 0.5 mg **54%**, 1.25 mg **60%** 3m CDP HR fingolimod 0.5 mg **0.70**, 1.25 mg **0.68**
FREEDOMS II[55]	Same as FREEDOMS	Same as FREEDOMS	1083 participants (ITT) 778 completed the trial age 41y disease duration 8.9y EDSS score 2.5	At 2y ARR reduction fingolimod 0.5 mg **48%**, 1.25 mg **50%** 3m CDP HR fingolimod 0.5 mg 0.83, 1.25 mg **0.72**
TRANSFORMS[56]	Same as FREEDOMS except for gap requirement after interferon beta or glatiramer acetate	1:1:1 weekly interferon beta-1a 30 µg (comparator) daily fingolimod 0.5 mg daily fingolimod 1.25 mg	1292 participants (ITT) 1153 completed the trial age 36y disease duration 6y EDSS score 2.0	At 1y ARR reduction fingolimod 0.5 mg **39%**, 1.25 mg **52%** 3m CDP HR fingolimod 0.5 mg 0.71

Table 1 *Overview of pivotal fingolimod trials*

Multi-country, pivotal, double-blind randomized controlled trials leading to marketing authorization of daily fingolimod 0.5 mg. Statistically significant results ($p<0.05$) are in **bold**. Data curated from trial reports and Gilenya summary of product characteristics.[53] RRMS: Relapsing-remitting multiple sclerosis, ITT: Intention-to-treat, y: year(s), m: month(s), EDSS: Expanded disability status scale, CDP: Confirmed disability progression, HR: Hazard ratio.

fingolimod's primary mechanism of action is redistribution of lymphocytes by retaining them in lymphoid organs, which results in a decrease in the amount of lymphocytes crossing to the CNS and hence a decrease in inflammatory activity.[54]

Fingolimod's marketing authorization was based on three phase III multi-site and multi-country RCTs, two of which had placebo as comparator whereas the third had an active comparator, interferon beta.[53] Initially, two of these trials (FREEDOMS and TRANSFORMS), were planned and commenced by the sponsor. They were followed by a third one (FREEDOMS II) as a result of a request by the FDA, so it took place mainly in the United States and addressed the issues raised by the FDA.[55] All three trials had similar eligibility criteria (**Table 1**). They included patients diagnosed with RRMS according to the McDonald 2005 criteria. The patients had to be aged between 18-55 years, have an EDSS score lower than 6.0, and have active disease demonstrated by at least one relapse during 1 year or two relapses during the 2 years prior to randomization. Those with an active infection, macular edema, immunosuppression, or any other clinically significant systemic disease were excluded.[54-56] Also excluded were those with a relapse or corticosteroid treatment during 30 days before baseline or natalizumab use during 6 months before baseline. An additional requirement for enrollment to the placebo-controlled studies were at least 3 months of gap after interferon beta or glatiramer acetate use.[54,55]

FREEDOMS and FREEDOMS II were double-blind three-arm trials in which the patients were randomized to receive daily placebo, fingolimod 0.5 mg, or fingolimod 1.25 mg (1:1:1) for 2 years. Following baseline, scheduled visits and neurological examinations occurred at 2 weeks,1 month, 2 months, 3 months, and every 3 months afterwards. MRI scans were taken at baseline and at months 6, 12, and 24 following baseline. The primary hypothesis was that compared to placebo, those on fingolimod (separately for both doses) had a lower ARR at 24 months in the intention-to-treat (ITT) population. This was tested by a negative binomial model of the number of relapses explained by the treatment arm, adjusting for logarithm of time under study (by an offset term), country, number of relapses in the 2 years prior to the study, and EDSS score at baseline. As key secondary endpoints, time-to disability progression confirmed at 3 or 6 months (CDP) between treatment arms were compared with log-rank tests and Cox proportional hazards (PH) regressions adjusting for country, age, and EDSS score at baseline. Time-to relapse was also analyzed with Cox PH regression as a secondary analysis. Time-to relapse and CDP were also described with Kaplan-Meier curves stratified by treatment arm. Secondary endpoints based on MRI scans included being free of Gd-enhanced T1 lesions or being free of new or enlarging T2 lesions at 6, 12, and 24 months and were analyzed with adjusted logistic regressions. Safety data were only described in frequencies per treatment arm and not statistically tested. Statistical tests were performed with an alpha value of 0.05 in a hierarchical testing framework to account for the multiplicity of the hypotheses.[54,55]

In FREEDOMS, 1272 patients were randomized, of which 1033 completed the study. These numbers were respectively 1083 and 778 in FREEDOMS II. In the FREEDOMS trial, the ARRs were significantly lower, 54% and 60% relative reductions respectively, for the 0.5 mg and 1.25 mg daily dosing of fingolimod compared to placebo. Similarly, in the FREEDOMS II trial, the respective relative reductions in ARRs were 48% and 50% in the fingolimod arms. In both studies, times-to first relapse were significantly longer in the fingolimod arms than in the placebo arms. In terms of secondary endpoints, the risk of 3-month CDP was lower in fingolimod arms significantly in the FREEDOMS trial (hazard ratio (HR) 0.70 and 0.68 for 0.5 mg and 1.25 mg) but non- or borderline significantly in FREEDOMS II (HR 0.83 and 0.72). In both trials, all tested inflammatory endpoints based on MRI scans were significantly lower in participants receiving fingolimod than those receiving placebo.[54,55] In the FREEDOMS trial report[54], the number of adverse events are evaluated to be close among treatment arms, except for

slightly higher risk of lower respiratory tract infections, bradycardia, macular edema, and increase of alanine aminotransferase in fingolimod arms. In FREEDOMS II[55], respiratory tract infections, herpes infections, and basal cell carcinoma risks were slightly increased in fingolimod arms. Hypertension, increased liver enzymes, and lymphopenia were other adverse events reported slightly more frequently for participants receiving fingolimod.

In the three-arm phase III RCT with active comparator, TRANSFORMS, patients were randomized to daily fingolimod 0.5 mg, 1.25 mg, or weekly intramuscular interferon beta-1a 30 µg (1:1:1) for a period of 12 months.[56] Endpoints were similar to those in the FREEDOMS trials except for the observation timespan. A total of 1292 patients were randomized, of which 1153 completed the study. Fingolimod, at both doses, was significantly superior to interferon beta-1a in reducing the ARR, elongating time-to first relapse, reducing the number of new or enlarging T2 lesions and Gd-enhanced T1 lesions. There was no significant difference between the treatment arms in terms of 3-month CDP. In terms of safety, bradycardia and atrioventricular block were the most commonly reported SAEs, which were common in the 1.25 mg fingolimod arm. In the TRANSFORMS trial report[56], four deaths in the 1.25 mg fingolimod arm are mentioned: two during the study respectively due to viral infections of varicella zoster and herpes simplex encephalitis, and two after the study respectively due to pneumonia and metastatic breast cancer.

Due to similar efficacy observed with the investigated doses of fingolimod daily 1.25 mg and 0.5 mg, but a dose-dependent increase in reported adverse events, fingolimod eventually received marketing authorization only for the lower dose of daily 0.5 mg. As expected from drugs with immunosuppressant mode of action[47,57], infections and neoplasms are listed among fingolimod's safety risks and warnings.[53] A Cochrane review synthesizing the results from six RCTs until February 2016 concluded that there was moderate evidence on efficacy of fingolimod in reducing the inflammatory disease activity compared to placebo, as measured by relapses and MRI, but there was little to no evidence on its efficacy in prevention of disability progression.[58]

In a systematic review on the infection risk of fingolimod, the pooled results from 12 RCTs showed a significant and robust risk increase (16%) in those receiving fingolimod as compared to placebo or an active comparator (one of interferon beta, glatiramer acetate, or natalizumab).[59] In subcategories of infection, the risk was significantly high for lower respiratory tract and herpes virus infections. Incidence of cancer on fingolimod has been meta-analyzed.[60] However, its result is less interpretable because a long time span is required to observe cancer cases, the reported measures were not based on a comparator group but were rather absolute, and both cohort studies and RCTs were included. Observational studies on the risk of cancer with fingolimod use have reported conflicting results.[61-63]

For the subgroup analyses of fingolimod, the groups were pre-specified based on baseline demographic variables, treatment history, and disease characteristics. Additionally, five patient groups were defined post-hoc by a combination of patients' disease activity and previous treatment history.[64] The same clinical efficacy outcomes (ARR and 3-month CDP) and the same modeling methods (respectively negative binomial and Cox PH) as in the main trial reports were used in the subgroup analyses. However, with the intention to address the small sample size of some subgroups and possible collinearity between subgroup variables and adjustment variables, the variables adjusted for were reduced in the subgroup analyses compared to the main analysis. In the subgroup analyses of the FREEDOMS trial daily fingolimod 0.5 mg treatment was found to have the greatest ARR reduction in those younger than 40 years-old and in men. The greatest and statistically significant reductions in risk of 3-month CDP were observed in men, treatment-naïve patients and in patients with a more severe course of disease at baseline in terms of higher EDSS, higher number of relapses, and higher T2 lesion volume.[64] In the pooled subgroup analysis of FREEDOMS and FREEDOMS II trials, the factors that

interact with fingolimod's effect on ARR reduction were in line with those from the FREEDOMS trial only. Subgroups with more active disease (measured by number of relapses or Gd-enhanced T1 lesions), with a higher disability (measured by EDSS), or with previous treatments at baseline had higher than average ARRs.[65]

1.2 Prognostic and treatment response prediction

1.2.1 Definition and significance

Prognosis is the possible outcome of a health condition that can be observed over time.[66] The outcome in question is usually binary as a future disease state, but may also be numeric. Although the average prognosis or the average treatment effect in a patient population may be of interest, more informative and useful is taking patient characteristics into account to derive individualized probabilistic predictions.[67] As a pillar of personalized medicine, the main question of interest in a prognostic study is "What is the risk of an outcome over a specified period of time based on individual characteristics at baseline?".[68,69]

Individual variables can be investigated to determine if they (independently) predict a future health outcome, i.e. if they are prognostic. The simplest prognostic prediction model can be formed by stratifying the patient population into two groups by the value of a single prognostic variable being below or above a threshold and assigning absolute event (1.0) and improbable event (0.0) probabilities to resulting groups.[70] However, predicting prognosis as a combination of multiple covariates is expected to outperform single prognostic factors[71,72], especially in diseases with complex mechanisms like MS. An extension of the simple framework is combining multiple predictors with logical operators to predict absolute event classifications or counting multiple logical statements to create a gradient of risk. Many expert-based clinical prediction rules, including the (modified) Rio score in the MS field, are formed like this.[73] In comparison, statistical methods to prognostic modeling allow representation of a more complex and fine-tuned functional relationship from a combination of prognostic factors to the outcome.[74]

One of the main purposes of prognostic research is to inform the clinician and the patient about what to expect from the future of the disease, based on which treatment and life decisions can be made. Another goal of prognostic research is stratifying patients based on expected disease severity and forming subgroups for clinical trials or selecting which patients to prioritize in resource-limited settings.[75] Prognostic models can also be used as summary measures of confounder for adjustment in interventional or observational studies.[69,72] Prognosis research may also allow for discovery of new predictors or, as a by-product, may enhance the understanding of disease processes.[76]

Even though methods of explanatory and prognostic studies may overlap, the evaluation and interpretation of their results should be in line with the differences in their aims.[68,77] In prognostic research, one is interested in the accuracy of predictions for individuals and there is no concept of a "confounder" whereas in explanatory research one is interested in causal effect estimates of individual factors accounting for possible confounders.[68] Determining the best treatment course for a patient may rely only on the prognosis of the patient regardless of incorporating the effect of treatment in the prediction. Or it may rely on, as has been done traditionally, the marginal treatment effect observed in clinical trials representing the population of interest. But, a more comprehensive approach should take into account the individual characteristics alongside their effect on the heterogeneity in response to the treatments. Thus, individualized treatment response prediction is the intersection of prognostic prediction and counterfactual (causal) thinking for the treatment of interest.

During prediction of heterogeneous treatment effects, the question of interest becomes “What is the expected change (or difference) in risk with the new treatment compared to the standard (or no) treatment based (conditional) on individual characteristics at baseline (pre-treatment)?”.[67,78] The framework becomes counterfactual in the sense that the question requires prognostic prediction under both treatment regimens during the same timeframe, which is impossible to observe in a single individual.[75,79-81] Treatment response prediction is essentially a comparative question ideally requiring a randomized, or an unconfounded, framework. The simplest version of treatment response prediction is the traditional subgroup analysis in clinical trials. Such analyses look for patient characteristics according to which the effect of treatment varies. It should be noted that the goal of this thesis is prediction of treatment response in patients naïve to the treatment in question, which is different than the frameworks distinguishing responders and non-responders after the fact based on the observed response of patients that were on treatment.[73]

1.2.2 Methodology

A prognostic prediction model first and foremost requires predictability, an unknown possibility for accurate prognosis. Not all outcomes in the future can be prognosticated with a desired accuracy. This intrinsic predictability of an outcome is the upper limit of any prediction endeavor and cannot be overcome by scientific efforts. Also important is the prognostic value present in the available variables. Factors that occur after the time point of prognostication and very close to the outcome, or unknown factors may play a bigger role in the process, rendering an outcome less predictable. The candidate variables are somehow measured and the validity of such measurements also factors in to the success of a prognostic prediction model. The choice of prognostic predictors and their valid operationalization require previous scientific research and subject-area (medical) expertise to rely on. Finally, how well a chosen functional form can represent the real-life process of generating the outcome from the available predictors determines the success of predictions from a prognostic model. Only this final point is under the control of the modeler but even when the best model under the given conditions is found, the performance of the resulting model may or may not be deemed useful for a clinical purpose, which is to be judged by those considering to use it in clinical practice.[82]

In prognosis research, the question of “when” becomes critical. The time point of prognostication and the time horizon of outcome determination need to be clearly defined. In this longitudinal framework, a prospective cohort study is the ideal design. Prognostic research has been categorized into the following consecutive phases: 1) Development 2) Validation (and updating) 3) Impact 4) Implementation.[68,74,83] The bulk of the literature is composed of development studies, and a small proportion is comprised of validation studies. Impact studies are very rare. This has been the status quo for a long while.[83] All the first three phases of prognostic prediction modeling research are prerequisites to clinical implementation and the scarcity of validation and impact studies prevents realization of clinical benefit from this research field. In this thesis, development and validation phases are conducted and hence are explained in dedicated subsections below.

Once the predictions from a developed prognostic model are found to be sufficiently valid for its intended use across different populations, an impact study – a comparative and usually a randomized interventional endeavor – is conducted to evaluate whether the use of the prognostic model during clinical decision making actually adds value to routine care by improving disease area relevant patient outcomes or reducing care costs.[72,74] Even when its positive and clinically relevant impact can be demonstrated, widespread implementation and uptake of a clinical prediction model into clinical practice require not only user-friendly interface for model application but also ease of predictor collection.[83]

1.2.2.1 Development

Development of a prediction model encompasses preprocessing of the data, including dealing with missing, and identification or selection of important predictors. Ideally, the candidate predictors should be selected via domain knowledge and be limited in number, especially for small sample sizes. Yet, statistical criteria and methods for variable selection, or modeling methods that intrinsically select variables can be employed. Using predictors (and outcomes) with standard definitions and objective measurements reduces the variance during the model development process, increasing the chances of generalizability. Also, predictors should precede the intended time point of prognostication. The outcome and its timing should be clinically significant, relevant to patients, and well-defined. Any variable that precedes the outcome in a causal graph, either directly or as surrogates of causal variables, are candidates to consider.[71] For facilitating adaptation by the healthcare providers and financiers alike, predictors that are easy and cheap to measure should be preferred.[74]

The development of a prediction model is the process of establishing a function from the selected predictors to the target outcome, which tends to be binary in prognostic research. Traditional methods used for this purpose have been logistic regression and Cox PH regression, depending on whether the event status or time-to-event is of interest.[84] In such a scenario, the model development process is about finding the coefficients of the predictors and establishing the baseline risk. Recently, machine learning methods like random forest or support vector machine have gained popularity.[85] These methods usually make less assumptions but are more data hungry due to this freedom. Yet, the models developed by traditional methods are more interpretable. The simpler and easier to interpret a prognostic model, the higher its chance of being accepted by healthcare providers and being used in the clinical practice.[86]

Data dependency of the modeling process tends to make the models overoptimistic within the dataset they are developed from. As noted before, model parsimony is desired: the less number of predictors the better.[86,87] However, there usually tends to be a high number of candidate predictors. This makes analytical variable selection, which itself is data dependent, desirable.[82] Any transformation of the predictors (e.g. interaction or higher order terms) uses degrees of freedom and increases the chance of overfitting to the development dataset. Another consideration that affects the overfitting of a model to a given dataset is the sample size.[88] The effective sample size is measured by the number of events relative to the number of considered variables (EPV), including all estimated parameters, i.e. the degrees of freedom.[89] It has been shown that the lower the EPV, the higher the overoptimism is, resulting in higher variability of estimated performance measures. The effect of EPV on performance measures is only slightly affected by varying the sample size and event rate combination.[90] The rule of thumb for a reliable model development is at least 10 events per considered predictor, although this blanket recommendation has been challenged.[89] Regardless, the main litmus test for a developed model is its evaluation in an independent dataset. If sufficient performance can be demonstrated in external validation, the way a model is developed becomes less relevant.[91]

It is recommended that prognostic models include the received treatment as a predictor, at least as main effect, which would assume a uniform effect of treatment across the population.[67,68,92] However, when the aim is to predict individualized treatment response, taking into account only the main effect would mean the same expected change in relative risk for treatment in all patients regardless of their individual characteristics. Thus, a model aiming to predict individualized treatment response includes interactions between the treatment and covariates to find how the treatment effect is heterogeneous by predictor values.[79,80]

The traditional subgroup analyses of RCTs use (ideally multiplicity-adjusted) univariate statistical tests to find out whether (ideally a limited number of) selected treatment covariate interactions are significant.

Statistical significance is taken to indicate a change in treatment effect in these subgroups.[88,93,94] Secondary analyses of RCTs for subgroup identification has the potential to have an inflated risk both for Type II errors due to low power of interaction detection and for Type I errors if many predictors are tested without taking into account the multiplicity. Also, the subgroups are usually formed with one variable at a time, or in a few narrowly-defined groups, followed by reporting of separate effect estimates for each. This is unlike how an individual patient is treated in the clinic where their many baseline characteristics may belong to different subgroups, maybe even with conflicting effect estimates, and an anticipated individual treatment response cannot be deduced from the existing subgroup analyses.[75,95] Also, subgroup analyses focus on maximizing the treatment effect heterogeneity, which does not necessarily translate to optimizing individual or population outcomes.[75] The uses, limitations, and possible dangers of traditional subgroup analyses have been extensively discussed elsewhere.[96-99]

In the spirit of the traditional subgroup analysis, treatment response prediction based on a combination of variables in the joint probability space has been presented as or intermingled with novel and more advanced methods of identifying subgroups.[79,80,100] For instance, a tree-based model can be used to perform the univariate split by also including higher-level interactions between variables but arrive at a subgroup based on the splits.[45] The goals of confirmatory, exploratory, descriptive, and predictive subgroup analyses should be differentiated.[87] Group-level or individual-level treatment effect heterogeneity can also be differentiated.[95] The data-driven discovery of the latter is the focus of this thesis and risk prediction modelling is used to approach it.

Treatment effect heterogeneity, even though not existing at the relative scale, can exist in the absolute scale, which is more relevant to a patient.[84,95] For instance, if the relative risk of experiencing a relapse within the next year under treatment A is 0.5 compared to placebo, this would translate to 40% absolute risk reduction for somebody with a baseline risk of 80% while it would mean 5% absolute risk reduction for somebody with a baseline risk of 10%. Thus, prognostic models identifying those under high risk might be useful for the decision to treat or not, even under the assumption of uniform relative treatment effect.[95,101,102] This approach to prediction of treatment effect heterogeneity has been called “risk-based method”.[103] Yet, models that also take into account the predictors that modify the treatment effect, if they exist, are of greater interest.[100] Prognostic factors and treatment effect modifiers are not necessarily the same set; however, an overlap is not unexpected.[71,101,104] Although conceptually useful, a strict distinction between prognostic predictors, which may modify the treatment effect at the absolute scale, and treatment response predictors, which are likely to be also prognostic, can be misleading. Both are relevant for evaluating the risk of individuals in the process of treatment decisions.[94,105]

Prognostic predictors have an effect on the risk of an outcome regardless of the treatment, whereas treatment response predictors modify the effect of treatment and are hence treatment specific. Using a single model simultaneously fit to treatment and control groups to find out the effect estimates for both prognostic and treatment response predictors (by including the main and treatment interaction terms) is the mainstream option to prediction of treatment effect heterogeneity.[79] This approach has been called “global outcome modeling”[45] or “treatment effect modeling”.[103] Alternatively, prognostic models can be constructed separately in the treatment and control groups, intrinsically taking into account the treatment interaction.[80] To predict treatment response for individuals, the risk of an event under new treatment and under standard treatment needs to be predicted in a counterfactual dataset - keeping all other covariates the same. The difference between these potential outcome predictions can then be considered the predicted benefit from the new treatment compared to the standard treatment.[45,93] Once this difference is calculated, others have gone further to model this predicted difference with a tree method (for example virtual twins[79]) or other non-parametric methods.[80] Other alternatives include focusing more on the treatment effect with interaction or model-based trees/forest. Another approach called “optimal

treatment regime" rather aims to optimize the binary treatment decision based on a model of treatment response[45,103] deeming the absolute risk or treatment effect irrelevant. Yet another proposed approach, called "local modelling", aims to model only the interactions, or the predictive part, omitting the prognostic component altogether.[45] Many novel statistical methods have been proposed for the problem of predictive subgroup identification. Yet, more methodological research is warranted in the field to determine recommendations for treatment response prediction modeling and methods for its evaluation.[45,78,103,106]

As argued by VanderWeele and colleagues[75], for optimizing treatment decisions, the method used and whether the final chosen model includes interaction terms are irrelevant as long as the model is good at predicting outcomes under alternative treatments conditional on the predictors. When the goal is treatment response prediction, the ideal study design is an RCT to ensure that there is a comparator and that treatment assignment is not confounded. Methods have been proposed in the causal inference literature to use observational data for this purpose[107], but they entail unverifiable assumptions.[81,94] The sample size requirement for the development of a treatment response prediction model is at least as much as a prognostic prediction model becuase it requires finding the difference in a pair of counterfactual outcome predictions. Identifying subgroups of patients with a combination of factors that would (not) benefit from treatment requires setting a threshold of how much difference from the marginal treatment effect is required to deem a subgroup or the observed heterogeneity clinically relevant.

1.2.2.2 Validation

Validation of a developed model is the evaluation of generalizability and transportability of its predictions to patient populations different than that is used in its development.[82,108] Underperformance in a validation sample is usually due to the differences between the development and validation samples caused by ommitting important predictors from the model or varying definitions and measurement methods in the samples. An alternative explanation to differences in validation performance can also be random variation, particularly when sample sizes are small.

In prognostic modeling, the goal is to optimize prediction for new observations, rather than existing data.[82,88,108] Thus, prevention of overfitting to the model development dataset becomes paramount. Evaluation of the prediction model in the training data, a.k.a. apparent validation, is likely to give overoptimistic results. Given a dataset, resampling methods exist which evaluate a prediction model's performance in a sample similar to the population of model development.[90] These may not only give unbiased estimates of the model performance but also rank different modeling methods[109] and tune the model complexity in partitions of previously unseen data. A basic method to assess the reproducibility of the predictive power via internal validation[110] is randomly splitting a proportion of the data (e.g. 2/3) for training the model and using the rest to test the performance of the model. However, this method is suboptimal in the sense that it reduces the effective sample size for both of the steps, increases the standard error of the results, and introduces a chance factor by a single split.[90,109,111] A modification to the random split framework splits the data into *k* parts (e.g. *k*=3) and uses the k^{th} part as test set while training the model in the rest. This repetition of the training/testing framework, a.k.a. *k*-fold cross-validation decreases the variability of the results and gives unbiased estimates with increasing *k*. Another more computer intensive method is using bootstrapping to develop the model in samples generated with replacement from the development set and evaluating the model in the main set or observations not included in the bootstrapped sample. Repetition of this bootstrapping procedure many times also gives unbiased performance estimates with low variability.[90] Although one can tune the parameters or get an unbiased performance estimate within a single dataset by cross-validation or bootstrapping, the final prediction model is generated by applying the chosen modeling method to the

whole development dataset using the optimal parameters and, if the modeling method does not have it intrinsically, using shrinkage methods to account for overoptimism.[90,109]

Internal validation based on resampling methods can be helpful during model development but is insufficient for providing a complete picture of a model's expected performance, especially if the developed model is intended to be used in different settings. Ideally, the aim of a prediction model is for it to be generalizable to patients other than it was developed in, demonstration of which requires evaluating the model in samples from a different setting, in terms of location, time, or data collection scheme. This is called external validation and is the ideal indicator of transportability of a prediction model. Confidence in a prediction model increases when its results are robust across diverse settings that are likely to have related populations.[110,112] Demonstrating good predictive performance in external validation is also recommended for treatment response predictions.[113] The performance of a prediction model is expected to be worse in external compared to internal validation.[74,114] A satisfactory performance of a prediction model in multiple external validation studies is sufficient for its success and when that is demonstrated, how the model was developed may eventually be considered irrelevant.[72,115]

The validation performance of a model can be assessed by different measures depending on the type of the outcome and the intention of the researchers. Goodness-of-fit measures commonly minimized in statistical modelling, like mean squared error for continuous outcomes or its equivalent Brier score for binary outcomes, can be used to give an overview of the model's performance when calculated in the validation dataset.[69,108] The Brier score, a strictly proper score[70,116], is the average squared distance between model predictions and the actual outcome for all observations. Because it measures loss, better predictions give lower Brier score and it is expected to be 0 for perfect model fit.[69,84] For choosing the best prediction model, the discrimination measure *c*-statistic, or its equivalent area under the receiver operating curve (AUC), is a natural choice. It evaluates how well the model distinguishes patients with high risk from those with low risk or, how well the rank of predictions correlates with the rank of those with or without the event.[90,112] *C*-statistic takes values between 0.5 to 1.0, respectively corresponding to no and perfect discrimination. As a rank statistic, *c*-statistic is independent of the scale of predictions, be it probabilities or any other score.

Another paramount measure of performance in prediction modeling is calibration, which is indicative of bias in predictions and measures how well the predicted probabilities match the actually observed probabilities.[108,112] Having a well-calibrated model is hence critical for its use in the clinical setting where treatment decisions depend not on the ranking of the patients but rather on the individual patient's risk of experiencing an outcome. Calibration is usually demonstrated with a graph in which the predictions are plotted against the observations and their agreement is assessed either by a line or in binned groups of patients. For a well-calibrated model, the line or points, respectively, are expected to lie close to the diagonal line with a slope of 1.[86] In the data that the model was developed, calibration is expected to be almost perfect by definition. The intercept (ideally 0) and the slope (ideally 1) of the line in the calibration plot further give information about how well a model predicts.[69,115]

If need be, a prediction model with a good discriminatory power can be recalibrated to fit the observations in a new dataset. However, no simple action can recover a model with low discriminatory power[109] and model revision is warranted by re-estimation or including new variables.[83] Hence, model and hyperparameter choices may depend only on discrimination in the development dataset but evaluation of the model quality should assess both discrimination and calibration in external validation. It should be noted that which performance measure values are acceptable or clinically significant remains to be determined by the medical experts for the disease-area in question and statistical criteria or *p*-values from tests (e.g. Hosmer-Lemeshow test for calibration) tend to be less relevant.[112]

A model that has good discrimination and calibration is not necessarily useful in the clinic if it does not perform better than blanket strategies at the risk threshold relevant for the decision making.[94] Let us assume a model predicting conversion from RRMS to SPMS within the next 5 years and which has a good AUC of 0.8 and good calibration. Let us also assume the model predicts the probability in the range of 30-60% and there is a moderately effective but very safe treatment option to prevent this conversion. Because conversion to SPMS in the short term, say with a prevalence 40%, is an outcome that physicians and patients would like to avoid, especially if there is a safe treatment, they would choose treatment even when the risk is only 10%. Then, the model is useless because one can simply treat every RRMS patient. However, if this treatment had serious side effects, then patients may be more reluctant to choose it and would require, say, 50% conversion risk to be treated with it. Then, the discriminative power of the model would be beneficial to sort out which patient has lower and which patient has higher risk than 50%. In the absence of such a high performing model, another model for the same outcome having an AUC of 0.6 and moderate calibration, can be useful if it can detect those patients with higher than 50% probability better than random decision to treat or not based on the prevalence. So, a model with moderate discriminative or overall performance may be very useful if the decisions based on it is beneficial in the risk range that is relevant for evaluating interventional options.[117]

Hence, to investigate the advantages of a prediction model in detecting true positives and true negatives, i.e. those who have the disease and those who do not, the sensitivity and specificity at a fine-grid of threshold points between the theoretical range of probabilistic cutoffs from 0% to 100% can be calculated. The sensitivity and specificity can then be combined with the event probability, i.e. the cutoff, to calculate the net benefit of the probabilistic model. Such net benefit can be visualized with plotting a decision curve in which the benefits of intervention to all or intervention to none are compared with intervention according to a prognostic score, generated by a model or a single biomarker, at varying risk thresholds.[118] An intervention may be a surgery, a drug, or any preventive measure that has its own benefits and costs. A risk threshold in the range of 0%-100%, or equivalently an event probability, at or above which an intervention would be desired but below which no action would be taken is chosen taking into account the potential benefits and harms of available interventions. Any decision-making based on probabilistic models first requires the decision of such a threshold. This choice should be done by area experts, preferably in consultation with patient groups. Then, the usefulness of blanket vs. model-based decisions can be compared at that risk threshold.[119]

The performance measures described above are useful for choosing, optimizing, and evaluating prognostic prediction models. Unfortunately, there are no established performance measures when the aim is to choose, optimize, and evaluate a scoring system for treatment response prediction and it is an active area of research.[113,120,121] The difficulty of this task arises from the fact that one cannot observe the effect of both treatments at the individual level due to the impossibility of observing counterfactual outcomes simultaneously.[81] Once calculated, the distribution of the expected treatment benefit per individual can be described by summary statistics or graphs for interpreting the extent of treatment effect heterogeneity.[93] *C*-for-benefit has been proposed as a novel measure for the evaluation of treatment response prediction scores, but it is limited in the sense that it assumes independence of outcomes under the alternative treatments given the score and has been argued not to be a proper scoring rule.[113,121] Thus, in this thesis, the focus is on optimizing and evaluating performance measures for the prognostic prediction of a model that contains both main terms and treatment interaction terms for selected baseline covariates. This decision entails the assumptions that such a model, if it performs well, is expected to also give reliable results when the difference in risk predictions under compared treatments is calculated, and that the set of variables considered are prognostic or predictive.[122]

1.2.3 Prognostic and treatment response prediction in multiple sclerosis

There is a great interest in prognostic prediction in MS due to the heterogeneity of the patient profiles, unpredictability of the disease course, and multiplicity of the treatment options.[20,49,123,124] Because RCTs only provide group-level effect estimates, an individualized approach with informative predictors is desired by healthcare professionals and patients.[20,125] In terms of treatment, the questions faced by the clinicians and patients alike are whether the patient needs to be treated or is expected to have a mild disease, to which treatment option the patient is likely to give the best response, and how much risk of experiencing a serious adverse reaction the patient has. At the time of diagnosis or treatment decision, a patient who is expected not to have a significant disability during the next few decades, even when untreated, should be distinguishable from those that will need a wheelchair in a few years.[20] Such concerns point towards the need to evaluate individual treatment decisions by balancing the benefits and risks.[49] Predicting the probability of response and risk of SAEs with available DMTs at the individual level would optimize treatment decisions and thus is an aspirational goal in MS.[44,49]

Although this interest in tailoring treatment decisions is sometimes framed as a biomarker discovery challenge, it is also a prognostic prediction challenge that requires application of appropriate study design and statistical methods. The lack of very strong predictors in MS makes prediction dependent on multivariable models.[49] The prognostic prediction literature in MS has focused on demographic variables (e.g. age, sex), clinical assessments of disability and functioning (e.g. Multiple Sclerosis Functional Composite (MSFC), EDSS), and history of clinical symptoms (e.g. time since onset, number of relapses in the previous year) as predictors of MS prognosis. With the passage of sufficient time to accumulate one, history of treatment with DMTs has also been a predictor domain that is considered in prognostic modeling studies. Cerebrospinal fluid (CSF) based biomarkers (e.g. oligoclonal bands, immunoglobulin quotients, neurofilament levels) have also been suggested as prognostic but their longitudinal investigations for prognostic prediction remain limited.[44] There are modeling instances based on omics (e.g. RNA, proteins) data, too, but finding a common thread among them to establish individual biomarkers is not possible. Comedications, concomitant diseases, non-CSF laboratory measurements (including that of vitamin D) and quality of life (QoL) measures have rarely been investigated in the prognostic modeling literature in MS even though they are considered to play a role in poor prognosis.[124]

In addition to their established acceptance as diagnostic markers of MS starting with McDonald criteria[17], measures derived from MRI have been extensively studied for their prognostic value and ability to predict treatment response.[19] Number of lesions and lesion volumes from T2-weighted and Gd-enhanced T1-weighted cranial images are the most commonly used and established biomarkers of prognosis early in the disease because they indicate demyelinating lesions.[126] Changes in these measures are considered as important markers of a patient's short and long-term prognosis throughout the disease, particularly for relapses.[44] The predictive power of many other MRI derived measures, like brain atrophy or lesions in the spinal cord, have been investigated and proposed by the research community but these are not yet widely accepted and there are barriers to their implementation into routine clinical practice.[19,44,127] Other types of imaging, like optical coherence tomography, and electrophysiological measurements have been studied but they are not investigated as widespread as MRI scans; so, more prognostic studies are warranted to demonstrate their value.[22,44]

Treatment response predictors and treatment effect modifiers that have been investigated in MS are similar to the prognostic predictors. Pre-treatment disease activity measured by relapses or MRI, and fast progression of disability are considered predictors of response to first-line DMTs. The identification of treatment response predictors is most advanced for the DMT marketed the earliest: interferon beta.[44,123] It should be noted that in the MS literature, biomarkers (or rules based on them) measured shortly after treatment initiation as surrogates of long-term treatment response have been incorrectly

termed treatment response prediction (e.g. Rio score[73]). These include post-treatment (e.g. 1-year) measurements of disease activity and neutralizing antibodies against interferon beta or natalizumab.[44] For methodological clarity, these are considered out of scope for this thesis. Sormani[128] explains the differences between prognosis, treatment effect modification, and surrogacy. In short, prognosis is independent of treatment, whereas treatment effect modification and surrogacy can only be discussed with reference to a specific treatment and a control group. Variables measured prior to the start of treatment are candidate treatment response predictors while surrogate markers of treatment efficacy or safety are measured post-treatment.

Due to infrequent external validations, high risk of bias during model development, considering non-routine and difficult-to-collect predictors in the models, and low compliance with reporting guidelines[22,23], the literature on prognostic prediction modeling in MS seems to be out of touch with the methodological requirements that would eventually lead to widespread clinical implementation of the prediction models. These are results from a systematic review on prognostic prediction models in MS.[129] The systematic review, which had a last database search date of July 2, 2021 and forward/backward search date of August 16, 2021, identified 75 models. Of these, only 12 had any external validation and two (Bergamaschi 2001[130] twice and Manouchehrinia 2019[131] thrice) had more than one external validation. However, the development and validations of all externally validated models had high risk of bias in at least one domain evaluated by the Prediction model Risk Of Bias ASsessment Tool (PROBAST[132]). Also, reporting quality was found to be poor. Of the 75 model developments and 15 external validations identified in the systematic review, only one model development[133] was evaluated to have low risk of bias in all domains and only two[133,134] were evaluated to have low risk of bias in the analysis domain of the PROBAST. Although neither of these prognostic prediction models were validated externally, it is worthwhile to note their properties and performance (**Table 2**).

In the prognostic prediction model development study that scored low for risk of bias in all domains of PROBAST, Pellegrini and colleagues[133] pooled the individual-level data from the placebo arms of four phase III RCTs of interferon beta (one trial), natalizumab (one trial), and dimethyl fumarate (two trials). In the multiply imputed datasets of more than 1582 participants with RRMS, who were followed-up for more than 2 years, the authors used competing methods of regularized regressions (separately lasso, ridge, and elastic net), support vector machines, and random forests (conditional and unconditional). The hyperparameters of the modeling methods were tuned via resampling methods. The outcome of interest was time-to disability progression defined as a composite of disability measured by EDSS, MSFC components (timed 25-foot walk test, 9-hole peg test (9HPT), and paced auditory serial addition test (PASAT)), or visual function test 2.5% (VFT). They included 23 baseline predictors in the domains of demographics, MRI, symptoms, disability, QoL, treatment, and adjusting factor of study identifiers. All the modeling methods they fitted had bootstrap-corrected survival *c*-statistic of less than 0.65. The authors evaluated this performance as poor and changed their strategy to using the abovementioned six methods to select the three most important predictors (PASAT, QoL physical component, and VFT). These important predictors were then used in an unpenalized Cox PH regression to generate the final model, which had even a lower bootstrap-corrected survival *c*-statistic (0.59, 95% confidence interval (CI) 0.57-0.61) at 2-years and a calibration slope of 0.97. The authors reported the HR coefficients from the model without the baseline hazard and concluded that prognosis of disability progression in MS was unpredictable with common clinical and demographical baseline characteristics.

In the other model development study that scored low for risk of bias in the analysis domain of PROBAST, De Brouwer and colleagues[134] used longitudinal data from an international MS registry (MSBase). After exclusion due to low quality and missing values, they analyzed data from a total of 6682

Study	Data source	Population	Outcome	Predictors	Modeling method[1]	Evaluation	Performance
Pellegrini 2019[133]	placebo arms of 4 RCTs	1582 participants with RRMS	Disability progression (time-to-event follow-up over 2y)	age, sex, ethnicity, Gd-enhanced T1 lesion number, T1 and T2 lesion volume, brain volume, brain parenchymal fraction, 1y and 3y pre-study number of relapses, disease duration, time since last relapse, EDSS, T25FW, 9HPT, PASAT, VFT, SF-36 physical and mental component, prior treatment	Predictor selection by regularized regressions, support vector machine, and random forests **Cox PH regression**	Internal validation bootstrapping	2y survival *c*-statistic (95% CI)[2] 0.59 (0.57-0.61) calibration slope 0.97
De Brouwer 2021[134,135]	MS registry (MSBase)	6682 participants with MS	6m CDP (2y)	age, sex, disease subtype, disease duration, number of relapses within 3y prior to baseline, EDSS at baseline and closest to 3y prior, maximum EDSS within 3y prior, difference between maximum and minimum EDSS within 3y prior, number of visits within 3y prior, last treatment, EDSS trajectories	Random forests, Bayesian tensor factorization, recurrent neural networks (time-aware and **continuous-time gated recurrent unit**)	Internal validation cross-validation	AUC (95% CI) [2] 0.66 (0.64-0.68)
Chalkou 2021[136]	MS registry (Swedish)	1752 2-year periods / 935 participants with RRMS	Relapse (2y)	age, sex, EDSS, disease duration, months since last relapse, number of relapses 2y prior, prior MS treatment, number of Gd-enhanced T1 lesions	Bayesian mixed-effects logistic model	Internal validation bootstrapping	AUC 0.65 calibration slope 0.91
Kalincik 2017[137]	international MS registry (MSBase)	9193 participants with CIS and MS receiving interferon-beta, glatiramer acetate, fingolimod, natalizumab, mitoxantrone	Repeating events: 6m CDP, 6m confirmed disability regression, relapse Single events: conversion to SPMS, treatment discontinuation Continuous: change in the cumulative disease burden	age, sex, cerebral MRI, spinal MRI, 1st symptom, ARR, 1y pre-baseline number of relapses, on-treatment relapses, relapse phenotype, relapse type within the last 2y, relapses that affect daily living within the last 1y or 2y, severe relapse within the last 1y or 2y, relapses with poor recovery within the last 1y or 2y, EDSS, EDSS trajectory, EDSS change, functional system scores, number of prior treatments, time since last prior treatment, most recent prior treatment, most active prior treatment, CSF	Dimensionality reduction by generating principal components; 42 models: six on-treatment outcomes for subpopulations of seven treatments (Anderson-Gil) Cox PH for time-to-event outcomes and linear for the continuous	External validation in a separate registry (only accuracy) Internal validation random-split[3]	accuracy 79-96% for repeating events 31-47% for continuous 3-42% for single events *c*-statistic[3] (95% CI) relapse 0.56 (0.54-0.57) 6m CDP 0.63 (0.61-0.66) 6m confirmed disability regression 0.67 (0.63-71)
Bovis 2019[138]	laquinimod and placebo arms of 2 Phase III RCTs	1982 participants with RRMS	3m CDP (follow-up up to 2y)	age, sex, disease duration, number of relapses within 1y prior to baseline, EDSS, and presence of Gd-enhanced T1 lesions, T1 and T2 lesion volume, normalized brain volume	Cox PH regression fit separately to treatment and control arms, then taking difference in coefficients for score	External validation in a separate RCT (n=1456)	score by treatment interaction *p*-value $p<0.05$

Study	Data source	Population	Outcome	Predictors	Modeling method[1]	Evaluation	Performance
Pellegrini 2019[139]	dimethyl fumarate and placebo arms of an RCT	1123 participants with RRMS	ARR (follow-up over 2y)	age, sex, ethnicity, 1y pre-study number of relapses, disease duration, time since last relapse, EDSS, T25FW, 9HPT, PASAT, VFT, SF-36 physical and mental component prior treatment	**Unpenalized** and regularized (ridge, lasso, elastic net) negative binomial regressions fit separately to treatment and control arms, then taking difference in coefficients for score	External validation in a separate RCT (n=976)	score by treatment interaction *p*-value $p<0.05$
Stühler 2020[140]	Registry of neurology practices (NeuroTrans Data)	3433 participants with RRMS receiving dimethyl fumarate, fingolimod, glatiramer acetate, interferon beta, natalizumab, teriflunomide	ARR 3m CDP	age, sex, 1y pre-baseline number of relapses, disease duration, time since last relapse, disability, number of prior treatments, any prior second-line treatment, last treatment, duration of last treatment, an interaction term for last treatment with duration of last treatment, treatment initiated at baseline; further interaction terms of treatment initiated at baseline separately with any prior second-line treatment, gender, 1-year pre-baseline number of relapses, and disease duration	hierarchical Bayesian generalized linear models, negative binomial for ARR, logistic for CDP followed by propensity score-adjusted models with treatment term in patients that received the treatment predicted to have the greatest effect versus the patients that did not	Internal validation random-split	*c*-statistic ARR: 0.61 CDP: 0.55 beneficial, except for fingolimod, but mostly non-significant effects for the studied treatments,
Chalkou 2021[122]	natalizumab, dimethyl fumarate, glatiramer acetate, placebo arms of 3 RCTs and placebo arms of 9 RCTs	2000 participants with RRMS	Relapse (2y)	age, sex, ethnicity, region, weight, volume of Gd-enhanced T1 lesions, 1y pre-study number of relapses, disease duration, time since last relapse, EDSS, T25FW, 9HPT, PASAT, VFT, actual distance walked, SF-36 physical component and mental component, prior treatment	Risk models by logistic regression penalized with lasso or **unpenalized** followed by network meta-analysis for treatment effect and its interaction with the score	Internal validation bootstrapping	*c*-statistic 0.62 calibration slope 1.05 interaction terms small and not significantly different from null

Table 2 *Overview of multivariable prediction models for multiple sclerosis*

Noteworthy prognostic prediction (first three rows) and treatment effect prediction models for multiple sclerosis (MS). [1)]If multiple competing methods, the chosen one is in **bold**. [2)]Calculated from reported standard error. [3)]Reported in a separate publication.[141] RCT: Randomized controlled trial, RR: Relapsing-remitting, y: years, Gd: Gadolinium, EDSS: Expanded disability status scale, T25FW: Timed 25-foot walk, 9HPT: Nine-hole peg test, PASAT: Paced auditory serial addition test, VFT: Visual function test, SF-36: 36-item short-form health survey, PH: Proportional hazards, CI: Confidence interval, m: months, CDP: Confirmed disability progression, CIS: Clinically isolated syndrome, SP: Secondary progressive, CSF: Cerebrospinal fluid, ARR: Annualized relapse rate.

participants with MS. The following predictor domains were considered at an assigned baseline: demographics, disease subtype, symptoms, disability, treatment, and EDSS trajectories. As competing methods in a nested cross-validation setting, the authors used random forests (features from only baseline or all features), Bayesian tensor factorization, and recurrent neural networks (separately time-aware and continuous-time gated recurrent unit) to model CDP within 2-years based on changes in EDSS and confirmation within 6-months of the increase. All modeling methods had an AUC below 0.68. Although the authors presented the longitudinal methods as favorable, the difference in AUC between the methods utilizing all features or alternatively trajectories were at most 0.01. The AUC of the method of continuous-time gated recurrent unit recurrent neural network, the method initially favored by the authors, was 0.66 (95% CI 0.64-0.68).[135] The authors did not provide the final model and concluded that model performance becomes better when clinical patient history data is used in prognostic prediction of disability progression. It should be noted that an overlapping set of co-authors performed a related analysis with the data from the same registry, which is yet reported as a pre-print.[142] The quality of the work is unclear at this early stage. The main differences between the original and new analyses are inclusion of functional system scores as predictors, considering treatment and relapse trajectories as predictors, using multiple observations per patient, and performing a type of validation by leaving centers out. Results revealed the highest achieved AUC of 0.72 and important predictors of EDSS, functional system scores, and disease duration.

Another methodologically sound prognostic prediction model developed by Chalkou and colleagues[136] was published after the search period of the systematic review. From a cohort study (Swedish MS Cohort), they included 1752 2-year periods of 935 RRMS patients and applied multiple imputation to missing data. By a literature review, they selected predictors in the domains of demographics, disability, symptoms, treatment, and MRI to predict relapse as a binary outcome during a 2-year follow-up. They used a Bayesian mixed-effects logistic model with random intercept and slope terms for individual patients, who may have multiple 2-year observations included, and a Laplace prior distribution for shrinkage of the estimated effects. The performance estimates they reported were bootstrap-corrected for the optimism calculated via a fixed-effect model due to computational constraints. The corrected AUC from the internal validation was 0.65 and the corrected calibration slope was 0.91. The authors reported the coefficients for normalized predictors from their model for all the imputed datasets individually in addition to providing a web-application. They interpreted their results in the context of previously reported models predicting relapses that had AUCs in the range of 0.60-0.70 and suggested that the poor discriminatory performance of their model could point to the fact that reliable predictors for relapses may yet to be discovered. The authors also analyzed the benefit of the model with a decision curve, from which they concluded that the prediction model would be useful for decision-making when the event probability threshold for considering intervention is between 15 and 30%. Although this is a narrow range, probability of relapse predicted by the reported model was in this range for almost half of the patients in their study.

Methodology for prediction modeling of prognosis is much more established than that of treatment response.[72,78,94,101] This is also true in terms of the methodology to systematically review and critically appraise them. Hence, the status of this research topic in MS is more difficult to access and evaluate. Yet, there are notable works that had the objective to predict response to treatments other than fingolimod in the context of RCT (dimethyl fumarate[122,139], laquinimod[138], and natalizumab and glatiramer acetate[122]) and to multiple marketed treatments in registry datasets.[137,140]

Using a subset of the RCTs included in the prognostic modeling study by a similar author list[133], Pellegrini and colleagues[139] developed a treatment response prediction model. From one phase III RCT, they used the placebo arm compared to the pooled arms of twice and thrice daily frequency of dimethyl

fumarate to increase power, despite only twice daily is eventually approved and is available in the market. This dataset was used to develop the individualized treatment decision score model via cross-validation. The same treatment arms from another phase III RCT was used to externally validate the score. The authors did a complete case analysis of 1123 and 976 participants with RRMS, respectively in the development and validation datasets. The outcome of interest was ARR observed during follow-up of over 2 years and the considered baseline predictors similar to that in the prognostic modeling study[133] except the MRI domain. The performance criterion used for model (and variable) selection was the area under the AD(q) curve, which is the average treatment effect difference as a function of the quantile of patients who have predicted treatment effect difference less than c, a cut-off in the range of possible predictions. Minimization of AD(q) was expected to correspond to a better model due to the fact that low ARR ratios favored the treatment. The authors made use of a method[143] based on fitting prognostic regression models separately to the patients in the treatment and control arms and deriving an individualized treatment response score model as the difference in coefficients of the control arm from the treatment arm. Pellegrini and colleagues used a negative binomial link with observed time as offset in competing modeling methods of fully unpenalized and many regularized methods (ridge, lasso, elastic net), of which the hyperparameters were optimized by cross-validation. The best performing method was the unpenalized full model. This result is not surprising because when the maximum likelihood is penalized during modeling, the difference in treatment response predictions from a full model with treatment interactions fit to the study population is not equivalent to those from models fitted separately to the treatment arms.[102] To assess whether the score they derived is actually an effect modifier, the authors reported *p*-values from a model explaining the outcome with the treatment, the score, and their interaction in the external validation dataset. The interaction term was statistically significant ($p<0.05$) and the observed ARR reduction was significantly higher in high responders (25^{th} percentile) versus standard responders (75^{th} percentile) as predicted by the score. The authors evaluated important variables by a conditional random forest algorithm and reported good QoL physical component, young age, good visual function, no treatment history, and lower EDSS score as variables influential on greater treatment response. The authors reported the full model, discussed the unexpectedness of the important treatment response predictors, and presented their approach as proof-of-concept.

A team of authors[138] intersecting with that of Pellegrini and colleagues[139], used the same methodology[143] to derive an individualized treatment response score by optimizing AD(q) in three placebo-controlled Phase III RCTs of laquinimod, a drug that was discontinued at phase III of its development process for the RRMS indication and was never marketed. The outcome of interest was time-to 3-month CDP measured by increase in EDSS and the baseline predictor domains they considered in this complete case analysis were demographics, symptoms, disability, and MRI. They fitted a total of 511 Cox regressions with all the possible predictor combinations to the treatment arms in order to select variables in one RCT (training) of 1101 participants and chose the best models with scores that gave the lowest *p*-values for the treatment score interaction in a Cox PH model consecutively in the training RCT, test RCT of 881 participants, and their combination. The third RCT with 1456 participants was used for external validation. The final selected model that was externally validated revealed older age, female sex, lower number of relapses within 1 year prior to baseline, higher normalized brain volume, and presence of Gd-enhanced T1 lesions as predictors of better response to laquinimod in terms of disability progression. The treatment by score interaction term was statistically significant ($p<0.05$) in the external validation dataset. The authors reported the model coefficients alongside a constant to replace baseline hazard, and concluded by recommending their methodology be used in the trials of other approved drugs so that subgroups of responders can be identified.

Using three of the RCTs included in the prognostic modeling by Pellegrini and colleagues[139] in addition to placebo arms from nine different RCTs, Chalkou and colleagues[122] aimed to predict individualized treatment response. They used individual participant data in a network meta-analysis context and performed complete case analysis. Their method was based on risk modeling (i.e. main terms of predictors only) in a total of 2000 participants with RRMS receiving natalizumab, dimethyl fumarate, glatiramer acetate, or placebo. The outcome of interest was relapse, as binary, during a follow-up period of 2 years. They considered a total of 31 predictors in a lasso framework and 14 prespecified predictors in a full-model approach as two-alternatives to their baseline risk model with a logit link. The predictors that were selected by lasso or prespecified were from the domains of demographics, MRI, symptoms, disability, QoL, treatment. They internally validated the two developed risk models, which had very close bootstrap-corrected *c*-statistics, 0.62 for the full model and 0.60 for lasso. But, calibration of the full model was better demonstrated by a calibration slope of 1.05 as opposed to 1.54. When the individual risk scores generated from these models were used in fixed-effects network meta-analysis that includes treatment by score interaction, the effect of the three investigated treatments turned out to be statistically significant whereas the coefficients of the interaction terms were found to be small in effect and not significantly different from null. In this methods-oriented study, the authors reported that natalizumab had higher benefit than the other drugs in most risk groups except similar benefit in lower risk patients (less than 30% relapse risk). The authors reported all coefficients from the risk models and estimated coefficients for the treatment main term and treatment by score interaction term. They concluded that this RRMS-specific modeling approach was not ready for implementation in clinics without external validation. They also commented that even though the *c*-statistics of the risk models were low and the interaction terms were not significant, these were not necessarily limiting factors to detect treatment effect heterogeneity at the absolute scale.

Although the ideal setting for treatment response prediction requires a control group and preferably randomization to prevent confounding factors, two large studies that utilize registry data to predict response to multiple DMTs, including fingolimod, do so in a global outcome prediction framework and can be considered prognostic, at the very least. Kalincik and colleagues[137] used data of 9193 CIS and MS patients with complete minimum data from the MSBase registry. Subgroups by seven DMTs of interferon beta, glatiramer acetate, fingolimod, natalizumab, and mitoxantrone were formed and randomly split to 90% training and 10% test to develop the prediction models. The models were externally validated in 2945 patients from the Swedish MS Registry. Three time-to-event outcomes (over 6-month CDP, over 6-month confirmed disability regression measured by EDSS, and relapse) were conceptualized as repeating with an Anderson-Gil PH model. Two time-to-event outcomes (conversion to SPMS, and discontinuation of treatment) were considered terminal events in a Cox PH model. A linear model was used to represent the remaining continuous outcome of change in the cumulative disease burden, which was operationalized as the AUC of disability measured with EDSS. Without giving details to how or why, the authors reported that the attempt to model adverse events were unsuccessful. As a result, they had 42 models: six on-treatment outcomes for seven DMT subpopulations. They considered many predictors (over 70 degrees of freedom) measured prior to the treatment that the subgroup was formed of, in the domains of demographics, MRI, symptoms, disability, treatment, and CSF. For dimensionality reduction, principal component analysis was employed in the total dataset, and three components were created to include in prediction models separately in subgroups for each drug. Another two “adjustment components” were formed by the center, number of visits and EDSS measurements pre-treatment, and treatment start date. As evaluation, the authors reported accuracy at 4-years in test and external validation sets. Regardless of the treatment subgroup, the performance was reported to be similar in the test and external validation datasets. In the external validation dataset, the accuracy was evaluated to be very good (79-96%) for the repeating events, which had arbitrary

definitions for detection of the event, moderate (31-47%) for the continuous outcome of change in cumulative disability, and low (3-42%) for single events of conversion to SPMS and treatment discontinuation. The authors also evaluated important prognostic and treatment response predictors of disability progression based on results from univariate models within each subgroup.

Kalincik and colleagues[137] concluded that their models were useful and were going to be implemented into a software tool to aid physicians and patients in their decision-making. Yet, the way that they represented the predictions from the models by comparing predictions under different treatment conditions has been criticized by methodological experts.[144] The rationale behind the criticism was that between-treatment comparison is inherently biased in a non-randomized setting because the prediction models probably cannot adjust for the effect of known and unknown confounding factors in treatment assignment. Especially the use of separate models for the different treatments would make the output from the models incomparable. The critics argued against the use of a tool based on the study by Kalincik and colleagues.[137] Also problematic was separate baseline hazards between treatments due to the modeling in subgroups rather than the whole population which precludes any adjustment for confounding between treatment decisions.[145] An overlapping set of authors[141] went on to assess the added prognostic value from another predictor, the multiple sclerosis severity scale combining EDSS and disease duration, to three of their original treatment effect prediction models. This update study had its own methodological pitfalls, but interestingly *c*-statistics in a random split data were reported for the original models, which were moderate: 0.56 (95% CI 0.54-0.57) for relapse, 0.63 (95% CI 0.61-0.66) for CDP, and 0.67 (95% CI 0.63-71) for confirmed disability regression.

The other treatment response prediction modeling study based on registry data is also implemented into a software tool.[146] Stühler and colleagues[140] included 3433 RRMS patients with complete data, randomly split to 90% training and 10% test sets, from a German registry of neurology practices to predict treatment response under different drugs at the time of a treatment switch. The two on-treatment outcomes of interest were ARR, conceptualized as count in a negative binomial model, and over 3-month CDP measured by EDSS and conceptualized as binary in a logistic model. The included predictors, which were selected based on medical expertise, were in the domains of demographics, symptoms, disability, treatment, and adjustment factors of duration of treatment initiated at baseline as an offset and center as random intercept. Due to low number of patients on other DMTs, patients using one of the following six treatments were included: dimethyl fumarate, fingolimod, glatiramer acetate, interferon beta, natalizumab, and teriflunomide. The authors used hierarchical Bayesian generalized linear models with non-informative priors to develop what they called “prognostic” prediction models with the above mentioned predictors. In a more complex model they called “predictive”, they additionally included interaction terms with treatment initiated at baseline. The authors reported the eight most important terms and the direction of their effect, from the “predictive” models for both outcomes. Apart from the intercept, last treatment, the duration of last treatment, and their interaction terms were dominant in these lists. Also reported were the calibration plots for predictive models per outcome and per treatment, which they evaluated to be good for lower values but not as good for higher values. Cross-validated *c*-statistic in the training set revealed little difference between the performance of prognostic (CDP: 0.56, ARR: 0.65) and predictive (CDP: 0.58, ARR: 0.65) models, leading the authors to question whether treatment interaction terms were adding any predictive value at all. The performance of the predictive models in the random-split test set was worse (*c*-statistic for CDP: 0.55, ARR: 0.61). The authors also evaluated the predictive models in propensity score-adjusted generalized linear models with the treatment term in a subgroup of patients that received the treatment predicted to have the greatest effect on them versus the patients that did not. Although underpowered to detect a significant difference in many cases, the results overall showed beneficial effects for the studied DMTs, except for fingolimod. The authors concluded that the developed models were robust, accurate, and generalizable.

They suggested that these models would be updated with new predictors and incoming data regularly but warranted external validation. In another publication, an intersecting list of authors report comparable and improving discriminatory performance of their prediction models in updates of data from the same registry.[146]

Because Stühler and colleagues[140] used commonly available predictors and reported their methods relatively well, there was an initiative[147] to externally validate their models. However, this was not possible because the details of the full model were not reported in the publication and the authors did not want to share the model with outside parties. This was probably due to the fact that these models are trademarks and have been designated as a medical device.[146] Still, the authors were responsive about the details of their methods and rather than external validation, the replicability of their methods are currently being evaluated in an independent cohort by applying the same variable definitions and model development strategy in a similarly selected group of treated patients from OFSEP, a French MS registry.[148]

The primary motivation and funding for this thesis was an MS Use Case within a medical informatics project[149] that aimed to make routine care data available to researchers. The goals of this Use Case were to predict the disease course and treatment success and to identify biomarkers that allow individualizing treatment decisions.[150] To this end, a treatment decision score was developed[151] using the routine care data from Klinikum Rechts der Isar in collaboration with the physicians and methodologists in Technical University of Munich. A total of 475 adult CIS and RRMS patients, who were newly-diagnosed and treated for at most 6 months at the time of their first available T2 MR image (baseline) were included. A total of 65 predictors were used from the domains of demographics, symptoms, disability, MRI, CSF, non-CSF laboratory, and others (fatigue and depression). The target outcome was the probability of having no new or enlarging lesions in T2 MR images between month 6 and month 24 (treatment success) under no treatment or first-line treatment options, fingolimod not included. Approximately 60% and 40% of the participants received no or first-line treatment. Taking into account the irregularities in timing of consecutive images in the routine care, the model was developed with a transformation forest. The base for the transformation forest was an interval-censored time-to-event model with the independent variable of treatment as none or first-line. The cross-validated AUC was 0.62 and the top five important predictors were from the domains of MRI (presence of periventricular lesions, number of T2 MRI lesions), CSF (CSF-specific oligoclonal bands, IgA to albumin quotient), or symptoms (relapses from categories any other than numbness, paresis, optic neuritis, or neurological symptoms). The expected benefit from using the developed score was up to 20% increase in probability of treatment success. There is an ongoing multicenter prospective cohort study, ProVal-MS (German Clinical Trails Register[152] study ID: DRKS00014034), to externally validate this treatment decision score in a similar group of patients and the initial results from the external validation are expected to be available in 2024.

1.3 Current knowledge and the gap

RRMS is a debilitating disease, inflicting young adults. Its clinical manifestation is heterogeneous among patients and difficult to predict from onset. In addition, more than a dozen treatment options with varying safety and efficacy profiles complicate clinical decision making. Prognostic or treatment response prediction could benefit healthcare providers and patients alike. Multivariable modeling has been used for this purpose, but mostly with suboptimal methods. The few noteworthy prediction modeling studies point to the limits of our knowledge. When the outcomes of interest are related to efficacy (based on relapse, disability, or MRI), the current methodologically sound literature in RRMS patients suggests poor to moderate discriminatory performance (*c*-statistic around 0.55-0.70) of prognostic prediction

models developed using demographic, disability, relapse, and MRI-based predictors - regardless of the inclusion of treatment interaction terms or not. It is unclear whether there is treatment effect modification relevant for decision making or a prognostic model would be sufficient to address treatment effect heterogeneity. Treatment response prediction by multivariable modeling in RCT contexts has been employed for a handful of DMTs, which do not include fingolimod. Additionally, evaluation of whether safety-related outcomes can be predicted by multivariable models is a gap, addressing of which would be valuable for individualized treatment decisions.

2. Objectives

Given the calls for research on personalizing medicine in multiple sclerosis based on pooled clinical trial data and lack of established or commonly adapted predictive models to this end, we sought to predict prognosis and response to fingolimod by reusing data collected in Phase III randomized clinical trials.

2.1 Primary objective

The aim of this thesis was to develop, externally validate, and evaluate multivariable statistical models predicting response to fingolimod within 2 years of treatment initiation based on the predictors from various domains measured at study baseline. The primary endpoint of interest was time-to-first confirmed relapse.

2.2 Secondary objectives

Exploratory aims were:

- To develop, externally validate, and evaluate multivariable statistical models predicting other efficacy endpoints of CDP, and new or enlarging T2 lesions
- To develop, externally validate, and evaluate multivariable statistical models predicting safety endpoints of SAEs and treatment discontinuation, and infections and neoplasms
- To identify variables predictive of all the investigated efficacy and safety endpoints

3. Methods

This Chapter aims to report in detail the methods used to realize the objectives. Information on the design and population of the trials used as data source in this thesis are provided in Sections 3.1 and 3.2, respectively. The baseline predictors considered and the definition of the six outcomes targeted in the prediction model can be found in Sections 3.3 and 3.4. These are followed by details of statistical methods used for data description in Section 3.5.1. Missing imputation, modeling and optimization methods used in the prediction model development, as well as methods to assess important predictors are reported in Section 3.5.2. The statistical measures used for evaluating the prediction models via external validation are reported in Section 3.5.3.

3.1 Study design

For this thesis (referred to as "study"), datasets from two phase III double-blind placebo-controlled RCTs (referred to as "trial(s)") were repurposed (**Figure 3**). The dataset of the FREEDOMS trial[54] was used for model development, whereas the FREEDOMS II trial[55] was used for external validation. These trials are summarized in the Introduction. In short, the primary objective of both trials was to compare the relapse rate in RRMS patients under treatment with fingolimod or placebo for 24 months. Secondary endpoints of the trials included disability progression, lesions in MRI, and safety. Approvals from institutional review boards and patient informed consents were in place. Results from these trials and their conventional subgroup analyses are reported in detail elsewhere.[54,55,64,65]

The data from the FREEDOMS trials were made available to researchers by their sponsor, Novartis, via the data sharing platform Clinical Study Data Request.[153] To access the datasets, a research proposal was submitted to this platform in 2019 (Proposal Number: 11223) and was deemed appropriate after evaluation by the Independent Review Panel and the sponsor. Following the data sharing agreement between the sponsor and our research institution, the trial datasets were accessed in 2020 via the secure research environment of Clinical Trial Data Transparency System and all data manipulation and analysis took place within that system. Due to the anonymized nature of the data shared by the sponsor, this project was deemed exempt from ethics committee approval by the Ethics Committee of LMU Munich (Project Number: 19-838).

3.2 Study population

Participant recruitment to the FREEDOMS trial took place between January 2006 and August 2007 in 22 countries. Participants to the FREEDOMS II were recruited from 8 countries between June 2006 and March 2009. Respectively in FREEDOMS and FREEDOMS II, the ITT population comprised of 1272 and 1083 participants randomized to daily receive fingolimod 0.5 mg or 1.25 mg or placebo (1:1:1). The eligibility criteria of both trials were almost identical. The analysis population in this study was the ITT population, which includes all participants who were included after the screening visit, were randomized to and took at least one dose of study medication. In the analysis, the participants were grouped to the treatment they were assigned to, irrespective of what they actually received. For this study, included were only patients randomized to the arms with the approved dose of daily 0.5 mg fingolimod and placebo. Although using the total study population would have increased the power, as argued by

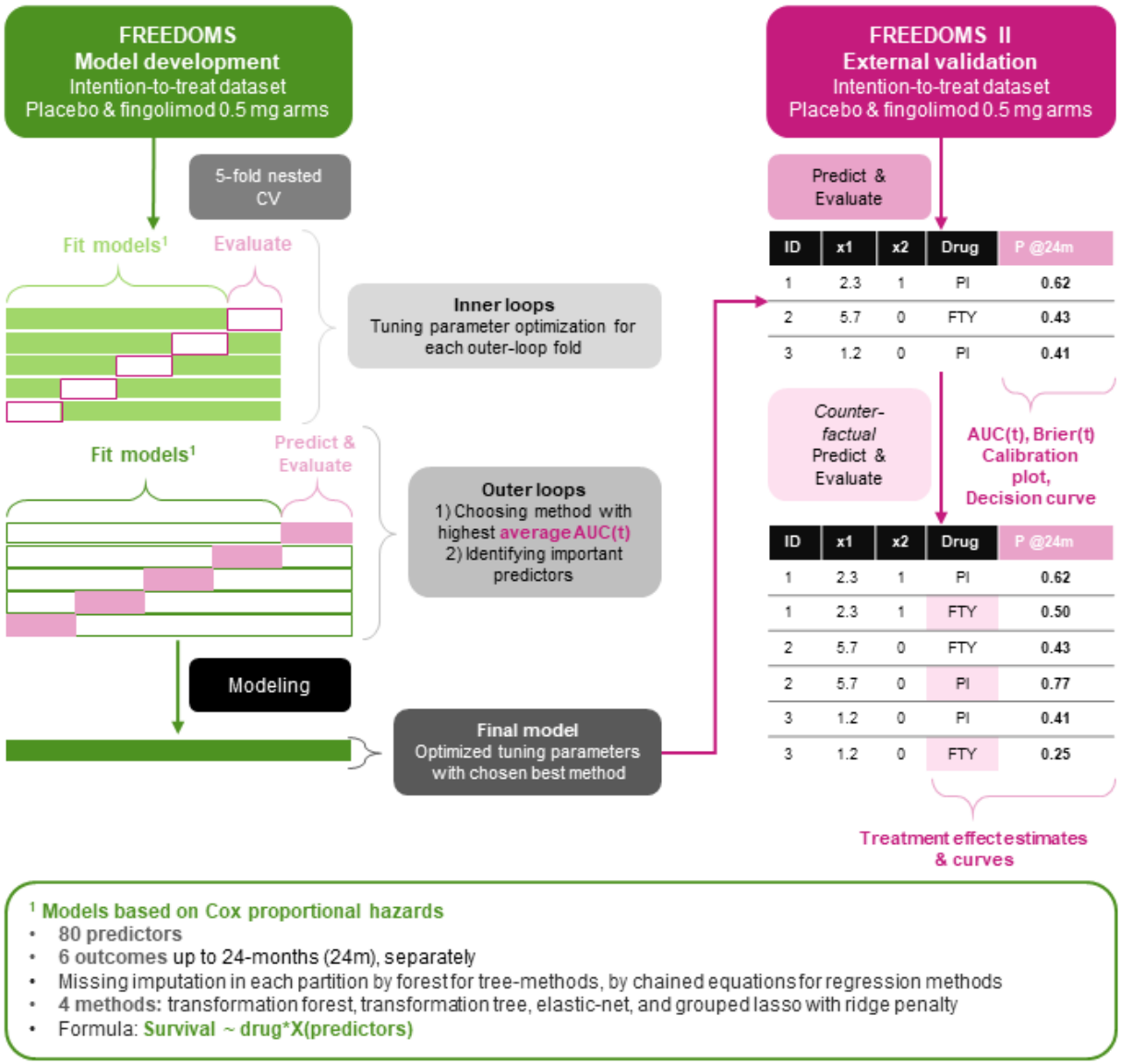


Figure 3 *Overview of methods*

Employed methods to develop and externally validate prediction models in the randomized controlled trial datasets of 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

Pellegrini and colleagues[139], the rationale behind this exclusion decision was that including the unapproved dosage in the model was irrelevant at best and erroneous at worst if the prediction model would be implemented in the clinic to make decisions regarding only the marketed dosage.

3.3 Predictors

The randomized treatment (drug) was conceptualized as a binary variable with categories of placebo or fingolimod (0.5 mg). As candidate predictors, a total of 80 variables were considered, in addition to the drug. These were collected at randomization (baseline) visit or, if a baseline value was missing, the screening visit, which, according to the trial protocol, took place earliest 45 days before baseline. The selection covered a wide range of domains from the common clinical (e.g. EDSS score, disease duration) or MRI parameters (e.g. number of Gd-enhanced T1 lesions) to those less investigated in MS prediction modeling, like comedications (grouped by ATC e.g. dermatologicals) or QoL measures

(dimensions of EQ-5D-3L[154]). **Table 3** gives an overview of the domains, number, and the list of variables considered. The comedication and concomitant disease variables with low number of positive participants were pooled to form combined variables. Except from age (eight categories of 5 year-windows from 16-55 years), all considered categorical variables had only two categories (yes/no) and were dummy-coded. Because the objective was to develop a treatment response model, the utilized modeling methods implicitly or explicitly included interaction terms of treatment with all the considered predictors. Including all levels and the interaction terms, the total number of terms considered in the regression models was 175.

3.4 Outcomes

Because the objective was to develop a treatment response model for prediction of multiple clinically relevant outcomes, the following events were defined.

Primary event:

1) Relapse: confirmed relapse, defined as in the source trial[54]

Secondary events:

2) New/enlarging lesions (T2 MRI): new or enlarging lesions in T2-weighted MRI scans measured during regular study visits
3) Confirmed disability progression (3m CDP): disability progression, as defined in the source trial[54], measured by EDSS during regular study visits and confirmed 3 months after its onset
4) Safety: safety outcome defined as a composite of SAE, discontinuation of the trial due to an adverse event, or death
5) Immunosuppressant safety (Immune safety): immunosuppressant-related safety outcome defined as an adverse event from the system organ classes (SOC) of infections and infestations or neoplasms, as coded by medical dictionary for regulatory activities (MedDRA)
6) Safety and efficacy (Composite): clinical efficacy and safety outcome defined as a composite of any of confirmed relapse (event 1), disability progression confirmed after 3 months (event 3), SAE, discontinuation of the trial due to adverse event, or death (event 4).

Many participants discontinued the source trials, leading to considerable amount of unobserved outcome values at 24 months (no event until day 765 and no visits between 676-765 days). Hence, to take different times of censoring into account, endpoints in this study were defined as time-to-first event since the randomization visit up to the 24-month visit. One month was considered to last 30 days. Although the desired time point of prediction was day 720, the prediction model was developed using the data available up to day 765 to make the predictions more stable.[84] The observations from participants without event were censored on the first of the trial participant's last visit day or day 765, which was considered to be the last acceptable day for a 24-month visit. The censoring in this study does not necessarily reflect the censoring definition in the source trial reports, which, for instance, have used a definition based on scheduled visits as a prerequisite of a censoring visit for time-to CDP.[64]

3.5 Statistical methods

The statistical analysis was performed in *R* version 4.2.0. The list of all packages used and their versions are available in Appendix A.

Domain	#	Predictors
Drug	1	Drug
Demographic	4	Age; Sex; Race; Body Mass Index (kg/m^2)
Clinical	13	EDSS score (total)
		EDSS functional system scores: Bowel and bladder; Brainstem; Cerebellar; Cerebral (or mental); Pyramidal; Sensory; Visual (or optic)
		MSFC: Mean of timed 25-foot walk; Mean of 9-hole peg test; Paced auditory serial addition test
		Visual acuity: Decimal score left; Decimal score right
Symptoms	3	Duration of MS since 1st symptom; Number of months since recent relapse; Number of relapses in the last 2 years
MS drug history	4	Number of prior MS treatments
		Prior DMT use: Glatiramer acetate; Interferon beta; Natalizumab or other MS treatment
MRI	4	Number of Gd-enhanced T1 lesions; Total volume of Gd-enhanced T1 lesions (mm^3); Total volume of T1 hypointense lesions (mm^3); Total volume of T2 lesions (mm^3)
QoL *EQ-5D-3L*	6	Anxiety/Depression; Mobility; Pain/Discomfort; Self-care; Usual activities; Visual analog scale
Comedications *classified by anatomical therapeutic chemical (ATC)*	11	Alimentary tract and metabolism; Blood and blood forming organs; Cardiovascular system; Dermatologicals; Genitourinary system and sex hormones; Systemic hormonal preparations, excluding sex hormones and insulins; Musculoskeletal system; Nervous system; Respiratory system; Various *Combined*: Antiinfective for systemic use or Antineoplastic and immunomodulating agents or Antiparasitic products, insecticides and repellents or Sensory organs
Concomitant diseases *classified by system organ from medical dictionary for regulatory activities (MedDRA)*	19	Congenital, familial and genetic disorders; Endocrine disorders; Eye disorders; Gastrointestinal disorders; General disorders and administration site conditions; Immune system disorders; Infections and infestations; Investigations; Metabolism and nutrition disorders; Musculoskeletal and connective tissue disorders; Neoplasms benign, malignant and unspecified (including cysts and polyps); Nervous system disorders; Psychiatric disorders; Renal and urinary disorders; Reproductive system and breast disorders; Respiratory, thoracic and mediastinal disorders; Skin and subcutaneous tissue disorders; Vascular disorders *Combined*: Blood and lymphatic system disorders or Cardiac disorders or Ear and labyrinth disorders or Hepatobiliary disorders or Injury, poisoning and procedural complications or Pregnancy, puerperium and perinatal conditions or Social circumstances or Surgical and medical procedures
Laboratory	16	*Hematology:* Absolute Basophils (10^9/L); Absolute Eosinophils (10^9/L); Absolute Lymphocytes (10^9/L); Absolute Monocytes (10^9/L); Absolute Neutrophils (10^9/L); Mean Cell Hemoglobin (fmol); Mean Cell Volume (fL); White Blood Cell (total, 10^9/L)
		Biochemistry: Albumin (g/L); Alkaline phosphatase (serum, U/L); Creatinine (µmol/L); Bilirubin (direct/conjugated, µmol/L); Gamma Glutamyltransferase (GGT), U/L); SGOT (AST, U/L); SGPT (ALT, U/L); Bilirubin (total, µmol/L)

Table 3 *Overview of candidate predictors*

The 80 predictors considered in the development of the prediction model.

3.5.1 Dataset description

The treatment arms in the model development and external validation datasets were separately summarized using median and (interquartile) range for continuous predictors, and by frequencies and proportions for categorical predictors. Also summarized was proportion of missing observations per predictor. The number of events were reported and the outcomes were described using Kaplan-Meier curves stratified by treatment groups. For sample size considerations during model development, events per variable were calculated by dividing the number of observed events to the total number of terms considered in regression models. Only for descriptive purposes, the events were considered as binary and the 24-month event status was plotted as presence of the event until last visit date or no recorded event but a visit between study days 676 and 765. The missing outcome was also summarized accordingly. The timing definition of 90 days around the 24-month visit supposed to happen on day 720 was chosen to reflect a conservative approach and is in line with the trials' visit definition for vital sign and laboratory measurements. It is not necessarily the same as the definition of study discontinuation in the original trial reports.

3.5.2 Model development

3.5.2.1 Modeling methods

Four modeling methods, all of which are based on Cox PH regression, were considered. The outcomes were conceptualized as time-to-first event rather than as binary or as count with an offset for time in study. This choice was motivated by the fact that the drop-out rate by month 24 was non-negligible in the trial reports. Of those randomized, 81% in FREEDOMS and 72% in FREEDOMS II were reported to have completed the study.[54,55] Distribution of reasons for study discontinuation were different in the active and placebo arms and a systematic review evaluated the source trials to be at risk of attrition bias.[58] Complete case analysis is not recommended in clinical prediction modeling[84] and imputation of the outcome by a certain method (e.g. based on a random forest or generalized linear models) would bring their own assumptions in the relationship of the outcome to the predictors and would thus interfere with the model optimization. Also, count models have the assumption of constant incidence rate across time, which does not necessarily reflect the observation of time-dependent ARRs in placebo arms of MS clinical trials.[155,156] The semi-parametric Cox PH model is very much related to logistic regression and Poisson regression[84] but does not require time-independent incidence rate, imputation of missing outcomes, or pre-specification of the shape of the time-dependent baseline risk.[157] One of its main assumptions is non-informative censoring, which means that censoring mechanism is independent of the outcome mechanism conditional on the covariates accounted for in the model. Also, a lack of predictor by time interaction terms entails an assumption of proportionality of the hazards over time, or, stated otherwise, time-constant effect of coefficients at the multiplicative scale. Like all generalized linear regression models, unless higher-order or interaction terms are included, Cox PH model has the assumptions of additivity and linearity of the predictor effects on the outcome at log-scale.[84] Another motivating factor in choosing time-to first event was the fact that these methods use the available information much more efficiently by taking into account also the time of event occurrence or censoring in addition to mere presence or absence.[157] For example, time-to first relapse was shown to have comparable power to ARR within conditions similar to that of the FREEDOMS trials and is considered to be a viable alternative endpoint.[158] The power of time-to-event analysis is expected to be higher in situations where the event rate is lower, like CDP, and binary outcomes tend to waste information.[157,159]

Two of the considered methods relied on recursive partitioning based on conditional transformation models (*R* packages *tram* and *trtf*). Conditional transformation models are a class of semi-parametric regression models that use transformation functions to allow the whole distribution of an independent variable be explained by predictors.[160] This gives the transformation models a capacity to represent and predict not just the mean, as in a traditional regression, but also higher moments of the outcome distribution.[161] Model-based recursive partitioning are tree methods in which the score function, closely related to a model's likelihood, is used as the splitting criterion. This allows the terminal nodes to be differentiated in terms of the model fit. In a time-to-event setting, this splitting criterion based on transformation models also allows deviations from the proportional hazards assumption.[162] When the regression model on which the tree is based contains explanatory variables, the tree that is formed can detect effect modification or subgroups of differential effect with respect to the splitting variables.[163] A transformation model-based tree can detect differing conditional distributions in its nodes. The limitations of a single tree are that interactions can only be represented in step structures by the splits, and the variability is high.[163,164] When trees are generalized to a random forest, smoother interactions can be represented, stability increases due to the regularization brought by randomness[165], and aggregation of many trees further reduces variability and brings stability.[105,166,167] Tree methods have the advantage of dealing with high-dimensionality and implicitly handling missing data by randomly assigning an observation with a missing value of the splitting variable to one of the children nodes with a probability of population distribution in the nodes.[168] In this study, the base model for tree methods was a Cox PH regression, baseline hazard of which was parameterized with degree five Bernstein polynomials.[162] This base transformation model contained treatment as the only explanatory variable so that the tree and the random forest would predict heterogeneity in treatment effect.

The other two considered methods were Cox PH regression models regularized with either an elastic net penalty[167,169] (*R* package *glmnet*) or a grouped lasso combined with ridge penalty[170] (*R* package *grpreg*). Regularizing regression models is recommended when the goal is prediction because it reduces the chance of overfitting by decreasing variance at the price of increasing bias.[77,84,167] Compared to ensemble methods, like random forest, the direction and extent of the predictors' influence on the outcome are easier to interpret with regularized regression methods. Yet, care is needed when interpreting the coefficient estimates in absolute terms because the regularization is expected to not only have introduced bias by shrinkage but also may have chosen one of the correlated predictors arbitrarily. Irrespective of whether the penalties have in-built variable selection (e.g. lasso) or not, regularized regressions are also able to deal well with high dimensionality even when the number of predictors is greater than the sample size and there is no unique solution to the ordinary likelihood estimation. The regularization of regression models is especially deemed important when treatment interaction terms are included in a prediction model, for which the conventional RCT is expected to be underpowered.[78] The elastic net penalty is a combination of lasso and ridge penalties, both of which were shown to be not superior than the other in all scenarios.[169] Elastic net ensures sparsity by predictor selection (characteristic of lasso penalty) while also ensuring that the coefficients of correlated predictors are shrunk comparably (characteristic of ridge penalty), introducing something similar to a grouping effect for correlated variables.[169] Grouped lasso applies a lasso penalty to predictor groups defined by the researcher and an additional ridge penalty can be introduced for further shrinkage of the individual predictors within the groups.[170] The dataset used in the regularized Cox PH regressions was formed of treatment, all predictors, and all possible treatment by first-degree predictor interactions. The main term for treatment was kept unpenalized to ensure its inclusion in the final model. The used *R* functions standardized the predictors before fitting the regularized regressions but their output were in the original scale. In the grouped lasso method, predictor main terms and their interaction with treatment were penalized together to ensure that they are both selected or dropped.[171]

The missing values in the datasets for the transformation model-based tree and forest were imputed by a random forest based method[172] (*R* package *missForest*), even though it was necessary for neither modeling nor predicting. The motivations behind this decision were making the comparison between tree methods and regularized Cox PH regressions comparable, and making the variable importance straightforward.[173] Missing imputation by iterative random forest fits to the complete data to predict the missing values in the predictors was used to create a single dataset. However, the fact that many trees are used to average the prediction can be considered to introduce multiplicity.[172] For the regularized Cox PH regressions, the missing data were imputed with predicted mean matching or logistic regression by chained equations[174] (*R* package *mice*), rather than a random forest based imputation method. Equation based missing imputation was chosen to ensure compatibility with the modeling methods and the future possibility of exporting the imputation method with fixed-chain equations[175] alongside the final prediction models. The dataset for the regularized regression with grouped lasso combined with ridge penalty was imputed once because the function could not handle weighting of the observations. The dataset for the regression with elastic net regularization was imputed 5 times. The multiple imputation was followed by stacking the imputed datasets. In order to take the uncertainty into account during modeling, a weight was assigned to each observation proportional to the amount of observed information divided by the number of imputations.[176] The outcome information was used during all imputations by including day of event or censoring and the Nelson-Aalen estimate of the cumulative hazard at that day in the dataset for imputation.[177] All imputations were performed separately in training, test, and external validation datasets.

3.5.2.2 Model optimization

In order to choose the best performing method and its parameters, nested *k*-fold cross-validation was applied in the model development dataset.[84,111,167] Folds were balanced in terms of the treatment arm (using *R* package *caret*). The tuning parameters of the competing models were optimized in the inner loops specific to the method within a training set of size n*(k-1)/k, where *k* was set to 5. The model optimized within and fitted to the training set was used for generating predictions in the remaining patients of the outer loop, i.e. the test set. The best model was chosen by evaluating the discrimination performance in the test set of each fold and comparing the average performance of the competing methods across folds. The discrimination was assessed by cumulative time-dependent AUC(t)[70,178-180] between baseline (day 0) and the average performance at three time-points: 6 months, 1 year, and 2 years, defined as days 180, 360, and 720. In order to check whether a performance measure that takes the actual predictions into account would give different results than the rank-statistic measure of AUC(t), the time-dependent Brier score, Brier(t), was also estimated for all models and time points as a sensitivity analysis.

Although the modelling method of choice was chosen by an overall discriminative measure in the test set of the outer loops, the parameter tuning was performed in the 5-fold inner-loops. For the model-based tree methods, parameters that maximized the log-likelihood in a 5-fold cross-validation within the training set were chosen. The tuning parameters of the tree controlled its depth. The considered alternatives were combinations of four values (between 0.05 and 0.20 with increments of 0.05) for the significance level for variable selection (alpha), and three values (between 5 and 15 with increments of 5) for the minimum acceptable number of observations at a terminal node. Hence, there were 12 possible combinations. The tuning parameters of the forest controlled the variability and the depth of its trees.[181] The considered alternatives were combinations of two values (square-root or one-third of the number of candidate predictors) for the number of predictors considered at each split, and the same three values as in the tree method for the minimum acceptable number of observations at the terminal

nodes of the trees forming the forest. The number of trees in the forest were set to 100. The function defaults (*trafotree* and *traforest*, respectively) were used for the remaining parameters.

For the tuning of regularized regressions, 5-fold cross-validation was implemented to find the alpha, mixing parameter of (grouped) lasso and ridge penalties, which minimized the error. There were 10 candidate alpha values between 0.1-1 with 0.1 increments, where alpha=0 imposes only a ridge penalty and alpha=1 imposes only a lasso penalty. The penalty parameter (lambda) that minimizes the error in a 5-fold cross-validation was chosen from 100 different values and was implemented by default by the *R* functions for the elastic net regression maximizing the partial likelihood and for the grouped lasso regression minimizing the deviance. The limited range of tuning parameter alternatives for random forest compared to those of regularized regression was not expected to pose a problem due to the low tunability of the random forest algorithm compared to that of elastic net.[182] A similar argument does not hold for the tree algorithm, which is expected to have higher variability compared to random forest.[167,182]

3.5.2.3 Variable importance

Separately for each outcome, the important variables from all the models developed in the training set of the outer cross-validation loops were recorded. Hence, for each modeling method and outcome combination, there were five sets of important variables. Any variable, as main term or as interacting with treatment, that was selected in or was considered important in at least two of the folds were considered important for that modeling method. Any variable that was found to be important for more than two modeling methods was considered to be important overall. Drug was not evaluated in this framework because it was either unpenalized in the regularized Cox PH regressions or integral to the base models of the transformation tree and forest.

The predictors selected via the lasso penalty, i.e. those with non-zero coefficients in the model developed in the training dataset, were considered important in the regularized Cox PH regressions. Similarly, the predictors selected by the tree as splitting variables in the training dataset were deemed important. In contrast to the other methods, the variable importance from the random forest was assessed in the test dataset. Based on the model developed in the training dataset, the mean of the difference in log-likelihood in the test set was calculated as is and after randomly permuting the individual predictors three times. Because the permutation-based importance is expected to vary around zero for non-informative predictors, the predictors that had importance greater than the absolute value of the (negative) minimum importance among the predictors were considered important.[166]

3.5.3 External validation

Separately for each outcome, the modeling method with the best average discrimination performance in the test sets of the outer loops of the nested cross-validation was chosen. The final model was generated by fitting the chosen modeling method to the whole development dataset. Final tuning parameters and the structure of the final model was described. The structure comprised of the baseline hazard and coefficients for regularized Cox PH regressions and a tree for the transformation model-based tree. A random forest is, unfortunately, not describable or neatly publishable. It can only be exported as a software object, which has the risk of compromising the privacy of the data that generated the model.[111]

Once formed, the final models were evaluated in the external validation dataset. The predictions in the external validation dataset were described by median and range. Cumulative AUC(t) and Brier(t), alongside their 95% CI, were plotted as a function of months since baseline using *R* package *riskRegression.*[70,183,184] Because the Brier score depends on the prevalence of the outcome, predictions

from a non-informative null model was used as a reference in the plot and in scaling the Brier score to take values between 0-100%. The higher the scaled Brier score is, the better the predictions are.[69,84] Also, receiver operating characteristic and calibration curves were plotted for all participants and by arm at three landmark times: 6 months (180 days), 12 months (360 days), and 24 months (720 days). Calibration curves were formed by creating 10 quantile bins for both arms, and individually for the treatment arms. The calibration-in-the-large was estimated by the coefficient of the intercept in a Poisson regression model of actual outcomes, adjusting for the expected number of events until censoring.[185] Calibration slope was estimated by the coefficient of the linear predictor in a Cox regression model of the actual outcome, adjusting for the baseline hazard.[84]

Also in the external validation dataset, the usefulness of the final prediction models was evaluated by decision curve analysis and treatment effect risk curves. Treatment effect predictions in the external validation dataset were described by median and range. With the decision curves, the net benefit of using the 24-month predictions from the model was compared to intervention to all or intervention to no patients for different decision thresholds representing the range of event risks at which treatment would be considered. The concept of intervention in this context is not necessarily fingolimod treatment, but rather any treatment regime, or preventive measure to decrease the risk of the specific outcome. Assuming that the default decision would be intervening with all the patients, the number of avoided interventions at different decision thresholds was visualized using the *R* package *dcurves*.[186] Then, expected treatment benefit, defined as decrease in the risk of outcomes, was estimated for all participants and was used to summarize the distribution of risk given treatment effect with the *R* package *TreatmentSelection.*[186] The extent to which there is treatment effect heterogeneity in response to fingolimod captured by the prediction model can be investigated by the range of treatment effect distribution and the risk given different treatments as a function of the population at increasing treatment effect quantiles. Like the number of avoided treatments, the treatment effect measures were calculated assuming the standard of care to be treating all. Also assumed was the threshold for model-based treatment decision to be any predicted benefit from fingolimod compared to placebo (i.e. treat if P(outcome under fingolimod) < P(under placebo)). The estimated treatment effect measures include proportion recommended treatment (fingolimod), empirical estimates[186,187] of average benefit of (no) treatment in those recommended (no) treatment, decrease in rate of outcomes under marker-based treatment compared to standard of care, variance in estimated treatment effect, and total gain defined as the integral of the difference between the treatment effect curve and the model-independent overall treatment effect.

4. Results

This Chapter follows the order laid out in Section 3.5, statistical methods. The sample size and event rates in the model development and validation datasets are described in Section 4.1.1. The population characteristics of the datasets are summarized and compared in Section 4.1.2. The cross-validated performance measures within the model development dataset and the final model for each outcome are reported in Section 4.2. The variables found to be important in outcome prediction are highlighted in Section 4.3. The external validation performance of the final models, evaluated first by discrimination and calibration, and then by decision and treatment response analyses, are reported in Sections 4.4.1 and 4.4.2, respectively.

4.1 Dataset description

4.1.1 Sample size and outcome description

The model development dataset was comprised of 843 participants, 425 of whom were in the active arm randomized to fingolimod 0.5 mg, and 418 of whom were in the control arm randomized to placebo. During follow-up, 331 (39%) participants in the model development dataset experienced a relapse, leading to a low EPV of 1.9 (**Table 4**), considering the degrees of freedom in the modeling process to be 175. Among all the outcomes, the number of participants that experienced an event (relative frequency, EPV) ranged from 119 (14%, 0.7) for the safety endpoint to 635 (75%, 3.6) for the immunosuppressant safety. The external validation dataset was comprised of 713 participants, 358 of

	Development			**External Validation**	
	***n* participants 843**	***n* fingolimod 425**	***n* terms 175**	***n* participants 713**	***n* fingolimod 358**
Outcome	**Events *(% participants)***	**Events Fingolimod *(% events)***	**EPV**	**Events *(% participants)***	**Events Fingolimod *(% events)***
Relapse	331 (39)	115 (35)	1.89	235 (33)	81 (34)
T2 MRI	525 (62)	207 (39)	3.00	417 (58)	168 (40)
3m CDP	166 (20)	72 (43)	0.95	170 (24)	80 (47)
Safety	119 (14)	54 (45)	0.68	128 (18)	72 (56)
Immune safety	635 (75)	319 (50)	3.63	570 (80)	295 (52)
Composite	469 (56)	192 (41)	2.68	405 (57)	188 (46)

Table 4 *Number of events*

Event frequencies in the model development and external validation datasets, overall and in the active arms of fingolimod 0.5 mg. EPV: Events per variable, number of variables based on total number of main effect and interaction terms considered in regression modeling. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

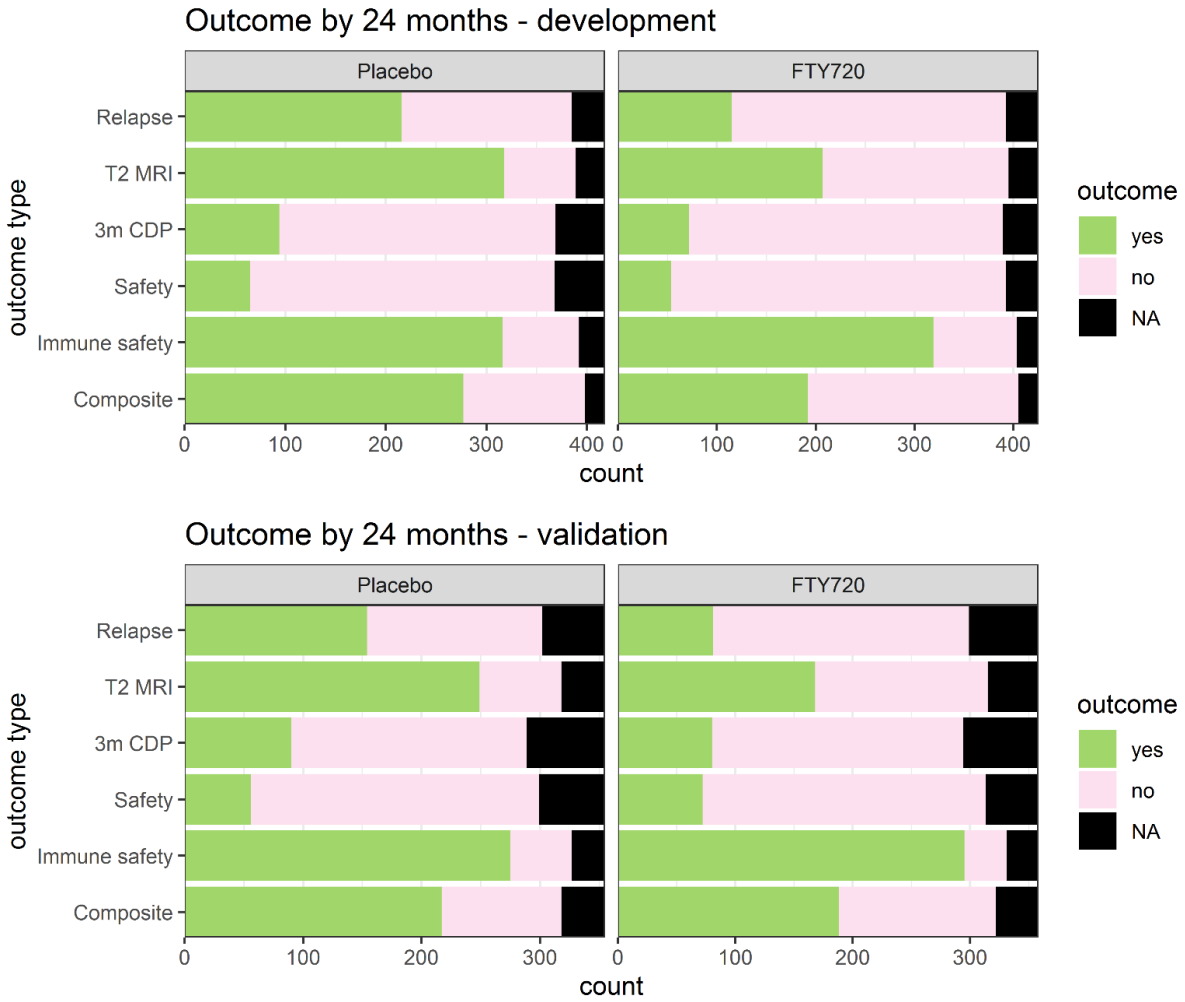


Figure 4 *Outcome frequencies*

Frequencies of outcomes conceptualized as binary at 24 months per trial arm. Fingolimod/placebo frequencies in the development dataset: 425/418, in the external validation dataset: 358/355. FTY720: fingolimod 0.5 mg, T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy; yes: event present, no: event absent, NA: missing outcome. *Upper figure*: Development dataset; *Lower figure*: External validation dataset.

whom were randomized to fingolimod 0.5 mg, and 355 of whom to placebo. The number of events in the external validation dataset were sufficient, i.e. above 100, for all outcomes. During follow-up, a total of 235 (33%) participants in the external validation dataset experienced a relapse. Similar to the model development dataset, the frequency of participants that experienced a safety event was the lowest at 128 (18%) and that experienced an immunosuppressant safety event was the highest at 570 (80%).

In the model development dataset, 112 (13%) participants did not have a 24-month visit defined as a visit between 676 and 765 days, whereas in the external validation dataset, 148 (21%) participants did not have a visit within this time-window. The visual description of the outcomes when conceptualized as binary at month 24 (**Figure 4**) reveal that the outcomes most inflicted by missing at month 24 are CDP and safety, due to low event numbers. Based on those events that were observed, the proportion of events in the fingolimod 0.5 mg arm to the total number of events ranged from 35% for relapse to 50% for immunosuppressant safety in the development dataset. In the external validation dataset, it ranged from 34% for relapse to 56% for the safety outcome.

The time-to-event outcomes used in this study are visualized by Kaplan-Meier curves stratified by arm (**Figure 5**). These crude graphs revealed that in the model development dataset, the probability of being event-free was most of the time higher in those treated with fingolimod 0.5 mg compared to those treated with placebo with respect to all outcomes, although the curves were not so well-separated for CDP and safety-related outcomes. Similarly, in the external validation dataset, fingolimod 0.5 mg looked superior to placebo for the outcomes of relapse and new or enlarging T2 MRI lesions. However, the curves were much less differentiated for the CDP outcome and being event-free seems somehow likelier in the placebo arm for safety-related outcomes. Another observation that the Kaplan-Meier curves hinted at was the pattern of visit frequency in the outcome of new or enlarging T2 MRI lesions. Because this event can only be observed during a visit with imaging, the event times, and hence steep drops in survival, are visibly concentrated around months 6, 12, and 24. In a subtler way, the Kaplan-Meier curve for CDP revealed similarly visible steps every 3 months.

4.1.2 Baseline description

Description of the participants by all predictors measured at baseline is provided in **Table 5**. Irrespective of the study or the arm, majority of the participants (over 70%) were female and in their 30s or 40s at baseline. Compared to the participants in the model development dataset, those in the external validation set were more likely to be female (81% / 77% vs. 71 / 70% in control / active arms), were slightly older with a longer disease duration (9.2 / 8.6 vs. 7 / 6.7 years), and were more than twice as likely to have used glatiramer acetate (41% / 36% vs. 11% / 10%) or interferon beta (59% / 61% vs. 28% / 30%) treatments prior to baseline. In all arms, the median EDSS score at baseline was 2 (interquartile range 1.5-3.5 in placebo arms and 1.5-3 in fingolimod 0.5 mg arms), the median number of relapses during the 2 years prior to baseline was 2 and the median time since recent relapse was about half a year. The participants in the development dataset had a higher load of T1 hypointense or T2 lesions in MRI but the participants in the external validation dataset had substantially more ongoing comedications and concomitant diseases in all groups defined respectively by ATC and MedDRA codes.

On average 0.3% (median 0%, range 0-6.5%) of the values were missing per predictor in the model development dataset. Although proportion of missing values per predictor was negligible, these were distributed over the trial population. In the model development dataset, 132 participants had at least one missing value, indicating that a complete case analysis would have excluded a considerable proportion (16%) of the participants. Only two predictors in the development dataset had proportion of missing values greater than 5%: the concomitant disease of general disorders and administration site conditions (26 (6%) / 29 (7%) in placebo/active arms) and albumin (21 (5%) / 24 (6%)). On average, 0.3% (median 0%, range 0-1.8%) of the values were missing per predictor in the external validation dataset. In the external validation dataset, 45 (6%) participants had at least one missing value. Of the 81 predictors, most had no missing at all in both the development (49 (60%)) and external validation (53 (65%)) datasets. The pattern of missing values is visually depicted in **Figure 6**. Values that are part of the same assessment (e.g. EDSS functional system scores) tended to be present or missing altogether for individual participants.

In the model development dataset, there were three participants with a relapse distance greater than 24 months and one participant with a relapse distance less than a month. In the external validation dataset, there were three participants with a relapse distance greater than 24 months and three participants with an EDSS score of 6 or 6.5. Participants with such protocol deviations were not excluded from this study as long as they were in the ITT group in the source trials.

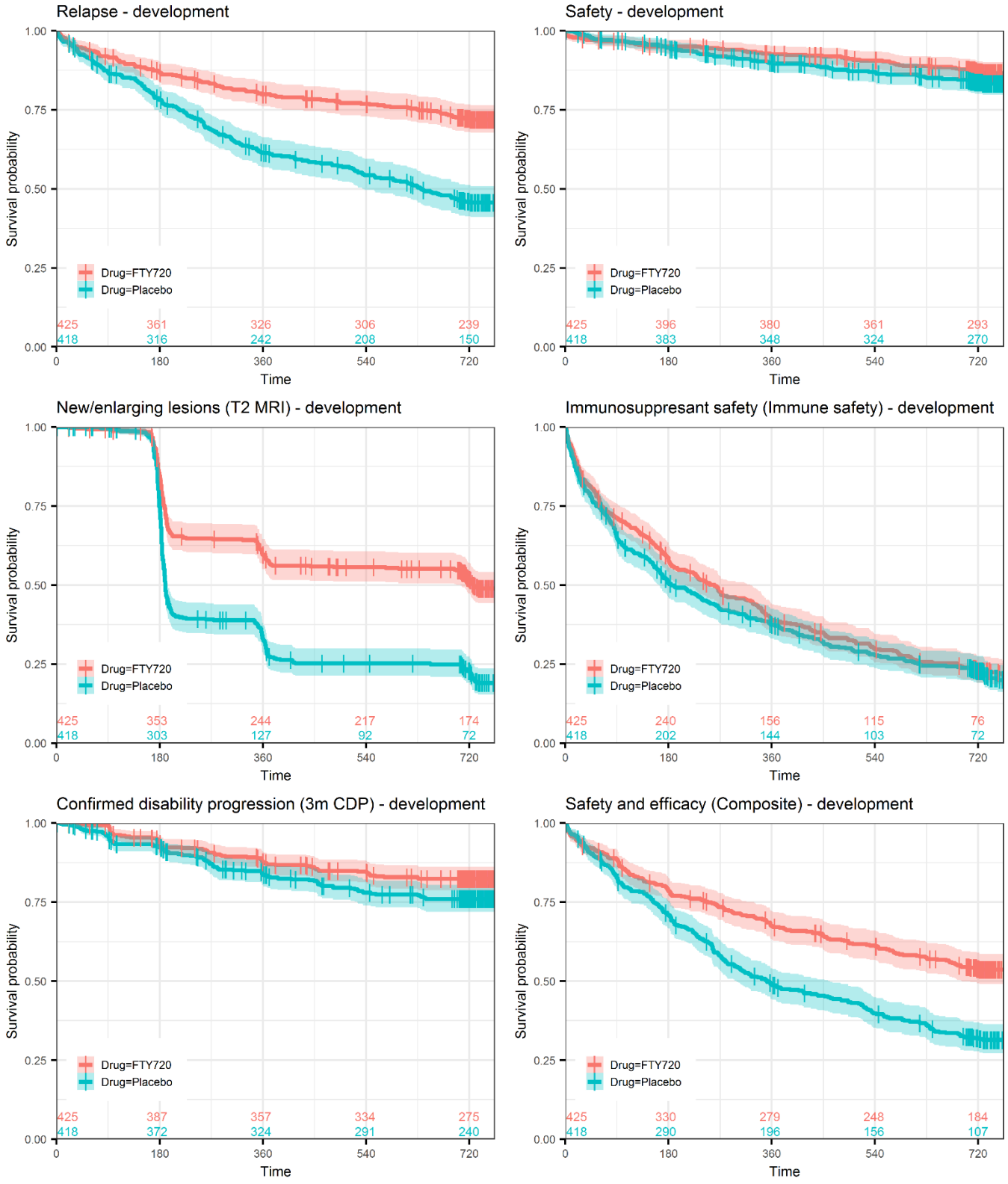
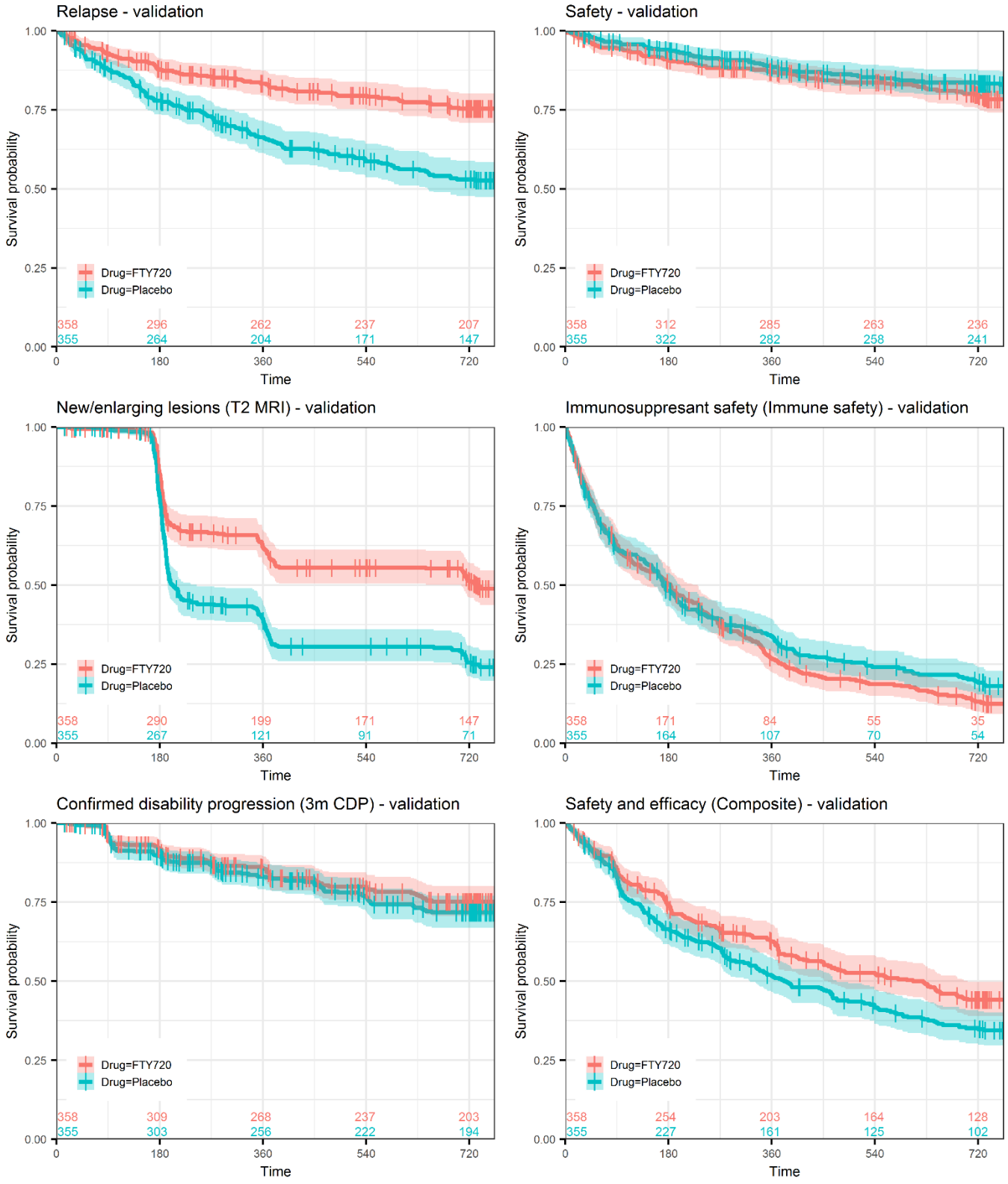


Figure 5 a/b *Kaplan-Meier curves*

Survival probability 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. **a)** *This page* Model development dataset **b)** *Next page* External validation dataset.


Relapse - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 296 262 237 207
355 264 204 171 147
0 180 360 540 720
Time
Safety - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 312 285 263 236
355 322 282 258 241
0 180 360 540 720
Time
New/enlarging lesions (T2 MRI) - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 290 199 171 147
355 267 121 91 71
0 180 360 540 720
Time
Immunosuppresant safety (Immune safety) - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 171 84 55 35
355 164 107 70 54
0 180 360 540 720
Time
Confirmed disability progression (3m CDP) - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 309 268 237 203
355 303 256 222 194
0 180 360 540 720
Time
Safety and efficacy (Composite) - validation
Survival probability
1.00
0.75
0.50
0.25
0.00
Drug=FTY720
Drug=Placebo
358 254 203 164 128
355 227 161 125 102
0 180 360 540 720
Time

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Demographic								
Age *21 - 25 (ref. 16 - 20)*	0 (0)	37 (9)	0 (0)	35 (8)	0 (0)	13 (4)	0 (0)	19 (5)
Age *26 - 30*		54 (13)		75 (18)		33 (9)		20 (6)
Age *31 - 35*		82 (20)		77 (18)		53 (15)		56 (16)
Age *36 - 40*		83 (20)		85 (20)		75 (21)		64 (18)
Age *41 - 45*		74 (18)		65 (15)		73 (21)		77 (22)
Age *46 - 50*		55 (13)		51 (12)		60 (17)		74 (21)
Age *51 - 55*		27 (6)		27 (6)		44 (12)		44 (12)
Body Mass Index *(kg/m^2)*	0 (0)	23.9 (21.4-27, 15.6-43.4)	0 (0)	24.1 (21.3-27.2, 17.2-49.1)	2 (1)	26.7 (22.7-31.3, 16.9-56.6)	0 (0)	27 (23.7-31.1, 13.9-50.8)
Race *non-Caucasian (ref. Caucasian)*	0 (0)	19 (5)	0 (0)	19 (4)	0 (0)	45 (13)	0 (0)	39 (11)
Sex *Female (ref. Male)*	0 (0)	298 (71)	0 (0)	296 (70)	0 (0)	288 (81)	0 (0)	275 (77)
Clinical								
EDSS score *(total)*	0 (0)	2 (1.5-3.5, 0-5.5)	0 (0)	2 (1.5-3, 0-5.5)	3 (1)	2 (1.5-3.5, 0-6)	3 (1)	2 (1.5-3, 0-6.5)
EDSS functional system scores								
Bowel and bladder	3 (1)	0 (0-1, 0-3)	2 (0)	0 (0-1, 0-3)	7 (2)	0 (0-1, 0-4)	6 (2)	0 (0-1, 0-3)
Brainstem	3 (1)	0 (0-1, 0-3)	2 (0)	0 (0-1, 0-3)	7 (2)	0 (0-1, 0-4)	6 (2)	0 (0-1, 0-4)
Cerebellar	3 (1)	1 (0-2, 0-4)	2 (0)	1 (0-2, 0-4)	7 (2)	1 (0-2, 0-3)	6 (2)	1 (0-1.2, 0-4)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Cerebral (or mental)	3 (1)	0 (0-1, 0-3)	2 (0)	0 (0-1, 0-3)	7 (2)	1 (0-2, 0-3)	6 (2)	0 (0-2, 0-3)
Pyramidal	3 (1)	1 (1-2, 0-5)	2 (0)	1 (1-2, 0-4)	7 (2)	1 (0-2, 0-4)	6 (2)	1 (0-2, 0-4)
Sensory	3 (1)	1 (0-2, 0-4)	2 (0)	1 (0-2, 0-3)	7 (2)	1 (0-2, 0-4)	6 (2)	1 (0-2, 0-5)
Visual (or optic)	3 (1)	0 (0-1, 0-3)	2 (0)	0 (0-1, 0-3)	7 (2)	0 (0-1, 0-3)	6 (2)	0 (0-1, 0-4)
MSFC								
Mean of timed 25-foot walk	2 (0)	5.1 (4.2-6.3, 2.5-91.5)	1 (0)	5 (4.2-6.2, 2.1-35.4)	5 (1)	5.2 (4.3-6.3, 2.5-46)	3 (1)	5.1 (4.4-6.2, 2.7-23)
Mean of 9-hole peg test	2 (0)	20.8 (18-24.3, 9.9-99.2)	2 (0)	20.4 (18.2-23.5, 11-67.4)	5 (1)	20.8 (18.7-23.9, 11.8-63.9)	3 (1)	20.8 (18.5-24.2, 9.3-64.6)
Paced auditory serial addition test	3 (1)	50 (42-56, 0-60)	2 (0)	52 (44-57, 4-60)	6 (2)	51 (41-56, 0-60)	5 (1)	51 (42-56, 0-60)
Visual acuity								
Decimal score left	7 (2)	1 (1-1, 0-1.7)	4 (1)	1 (1-1, 0-1.6)	3 (1)	1 (0.8-1, 0.2-1.5)	6 (2)	1 (0.8-1, 0.1-1.5)
Decimal score right	5 (1)	1 (1-1, 0-1.7)	0 (0)	1 (1-1, 0-1.6)	2 (1)	1 (0.8-1, 0.1-1.5)	3 (1)	1 (0.8-1, 0.1-1.5)
Symptoms								
Duration of MS since 1st symptom *(years)*	0 (0)	7 (3-12, 0.3-32.2)	0 (0)	6.6 (2.8-11.3, 0.3-34.9)	1 (0)	9.2 (4.9-15.2, 0.2-40.1)	0 (0)	8.6 (4-15, 0.2-49.1)
Number of months since recent relapse	0 (0)	5.2 (3.3-8.2, 1.2-62.9)	0 (0)	5.2 (3.5-8, 0.4-85.6)	1 (0)	5.7 (3.5-9.1, 1.4-28)	0 (0)	6 (3.6-9.4, 1.3-26)
Number of relapses in the last 2 years	0 (0)	2 (1-3, 1-10)	1 (0)	2 (1-3, 1-11)	1 (0)	2 (1-3, 1-14)	0 (0)	2 (1-3, 1-8)
MS drug history								
Number of prior MS treatments	0 (0)	0 (0-1, 0-4)	0 (0)	0 (0-1, 0-4)	0 (0)	1 (0-2, 0-5)	0 (0)	1 (0-2, 0-5)
Prior DMT use								
Glatiramer acetate	0 (0)	44 (11)	0 (0)	42 (10)	0 (0)	146 (41)	0 (0)	129 (36)
Interferon beta	0 (0)	115 (28)	0 (0)	127 (30)	0 (0)	209 (59)	0 (0)	218 (61)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Natalizumab or other MS treatment	0 (0)	54 (13)	0 (0)	50 (12)	0 (0)	48 (14)	0 (0)	41 (11)
MRI								
Number of Gd-enhanced T1 lesions	2 (0)	0 (0-1, 0-26)	1 (0)	0 (0-1, 0-84)	1 (0)	0 (0-1, 0-46)	1 (0)	0 (0-1, 0-33)
Total volume of Gd-enhanced T1 lesions *(mm^3)*	2 (0)	0 (0-101.5, 0-2970)	1 (0)	0 (0-82.7, 0-6849.8)	1 (0)	0 (0-77.3, 0-4060.1)	1 (0)	0 (0-94.4, 0-5570.3)
Total volume of T1 hypointense lesions *(mm^3)*	2 (0)	811.2 (205.9-2301.9, 0-20955.9)	1 (0)	814 (218.2-2402.1, 0-22377.8)	1 (0)	377.5 (75.8-1387, 0-17362.2)	1 (0)	343.4 (54.4-1293.4, 0-23937.3)
Total volume of T2 lesions *(mm^3)*	2 (0)	3416.2 (1291.8-8342.7, 0-37147.8)	1 (0)	3303.3 (1208.1-7895, 0-47147.6)	1 (0)	2702.4 (987.1-6996.5, 0-69202.6)	2 (1)	2356.2 (777.5-6123.1, 0-54369.4)
Quality of Life (EQ-5D-3L dimensions)								
Anxiety / Depression	2 (0)	1 (1-2, 1-3)	1 (0)	1 (1-2, 1-3)	3 (1)	1 (1-2, 1-3)	3 (1)	1 (1-2, 1-3)
Mobility	2 (0)	1 (1-2, 1-2)	1 (0)	1 (1-2, 1-2)	3 (1)	1 (1-2, 1-2)	4 (1)	1 (1-2, 1-2)
Pain / Discomfort	2 (0)	2 (1-2, 1-3)	1 (0)	1 (1-2, 1-3)	3 (1)	2 (1-2, 1-3)	3 (1)	2 (1-2, 1-3)
Self-care	2 (0)	1 (1-1, 1-2)	1 (0)	1 (1-1, 1-2)	3 (1)	1 (1-1, 1-2)	3 (1)	1 (1-1, 1-2)
Usual activities	2 (0)	1 (1-2, 1-3)	1 (0)	1 (1-2, 1-3)	3 (1)	1 (1-2, 1-3)	3 (1)	1 (1-2, 1-3)
Visual analog scale	3 (1)	79 (65-90, 24-100)	1 (0)	80 (70-90, 0-100)	4 (1)	80 (70-90, 20-100)	5 (1)	80 (70-90, 20-100)
Comedications								
Alimentary tract and metabolism	0 (0)	89 (21)	0 (0)	91 (21)	0 (0)	205 (58)	0 (0)	210 (59)
Blood and blood forming organs	0 (0)	15 (4)	0 (0)	18 (4)	0 (0)	78 (22)	0 (0)	73 (20)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Cardiovascular system	0 (0)	68 (16)	0 (0)	77 (18)	0 (0)	181 (51)	0 (0)	188 (53)
Dermatologicals	0 (0)	13 (3)	0 (0)	14 (3)	0 (0)	157 (44)	0 (0)	148 (41)
Genito urinary system and sex hormones	0 (0)	131 (31)	0 (0)	133 (31)	0 (0)	211 (59)	0 (0)	206 (58)
Systemic hormonal preparations, excluding sex hormones and insulins	0 (0)	18 (4)	0 (0)	10 (2)	0 (0)	30 (8)	0 (0)	42 (12)
Musculo-skeletal system	0 (0)	37 (9)	0 (0)	26 (6)	0 (0)	185 (52)	0 (0)	190 (53)
Nervous system	0 (0)	128 (31)	0 (0)	118 (28)	0 (0)	255 (72)	0 (0)	253 (71)
Respiratory system	0 (0)	19 (5)	0 (0)	22 (5)	0 (0)	106 (30)	0 (0)	115 (32)
Various	0 (0)	30 (7)	0 (0)	32 (8)	0 (0)	72 (20)	0 (0)	66 (18)
Antiinfective for systemic use or Antineoplastic and immunomodulating agents or Antiparasitic products, insecticides and repellents or Sensory organs	0 (0)	22 (5)	0 (0)	16 (4)	0 (0)	102 (29)	0 (0)	119 (33)
Concomitant diseases								
Congenital, familial and genetic disorders	0 (0)	13 (3)	0 (0)	15 (4)	0 (0)	16 (5)	0 (0)	19 (5)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Endocrine disorders	0 (0)	18 (4)	0 (0)	9 (2)	0 (0)	29 (8)	0 (0)	34 (9)
Eye disorders	0 (0)	42 (10)	1 (0)	55 (13)	0 (0)	88 (25)	0 (0)	76 (21)
Gastrointestinal disorders	0 (0)	33 (8)	0 (0)	30 (7)	0 (0)	102 (29)	0 (0)	96 (27)
General disorders and administration site conditions	26 (6)	13 (3)	29 (7)	15 (4)	2 (1)	30 (8)	1 (0)	36 (10)
Immune system disorders	0 (0)	46 (11)	0 (0)	44 (10)	0 (0)	151 (43)	0 (0)	158 (44)
Infections and infestations	0 (0)	34 (8)	1 (0)	37 (9)	0 (0)	80 (23)	0 (0)	75 (21)
Investigations	0 (0)	10 (2)	0 (0)	19 (4)	0 (0)	42 (12)	0 (0)	41 (11)
Metabolism and nutrition disorders	0 (0)	46 (11)	0 (0)	45 (11)	0 (0)	63 (18)	0 (0)	74 (21)
Musculoskeletal and connective tissue disorders	0 (0)	49 (12)	0 (0)	43 (10)	0 (0)	128 (36)	0 (0)	118 (33)
Neoplasms benign, malignant and unspecified (incl. cysts and polyps)	0 (0)	11 (3)	0 (0)	19 (4)	0 (0)	88 (25)	0 (0)	74 (21)
Nervous system disorders	0 (0)	87 (21)	0 (0)	91 (21)	0 (0)	208 (59)	0 (0)	184 (51)
Psychiatric disorders	1 (0)	73 (17)	1 (0)	69 (16)	0 (0)	187 (53)	0 (0)	176 (49)
Renal and urinary disorders	0 (0)	24 (6)	1 (0)	21 (5)	0 (0)	83 (23)	0 (0)	69 (19)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Reproductive system and breast disorders	0 (0)	12 (3)	0 (0)	20 (5)	0 (0)	47 (13)	0 (0)	49 (14)
Respiratory, thoracic and mediastinal disorders	0 (0)	19 (5)	0 (0)	26 (6)	0 (0)	70 (20)	0 (0)	70 (20)
Skin and subcutaneous tissue disorders	0 (0)	35 (8)	0 (0)	41 (10)	0 (0)	88 (25)	0 (0)	110 (31)
Vascular disorders	0 (0)	35 (8)	0 (0)	26 (6)	0 (0)	62 (17)	0 (0)	51 (14)
Blood and lymphatic system disorders or Cardiac disorders or Ear and labyrinth disorders or Hepatobiliary disorders or Injury, poisoning and procedural complications or Pregnancy, puerperium and perinatal conditions or Social circumstances or Surgical and medical procedures	0 (0)	34 (8)	0 (0)	45 (11)	0 (0)	113 (32)	0 (0)	110 (31)
Laboratory								
Hematology								

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Absolute Basophils *(10^9/L)*	0 (0)	0.1 (0-0.1, 0-0.3)	0 (0)	0.1 (0-0.1, 0-0.2)	0 (0)	0.1 (0-0.1, 0-0.2)	0 (0)	0.1 (0-0.1, 0-0.2)
Absolute Eosinophils *(10^9/L)*	0 (0)	0.1 (0.1-0.2, 0-0.7)	0 (0)	0.1 (0.1-0.2, 0-0.9)	0 (0)	0.1 (0.1-0.2, 0-0.6)	0 (0)	0.1 (0.1-0.2, 0-2.2)
Absolute Lymphocytes *(10^9/L)*	0 (0)	1.8 (1.4-2.1, 0.7-4.8)	1 (0)	1.8 (1.4-2.2, 0.6-6.2)	0 (0)	1.8 (1.4-2.2, 0.6-4.6)	0 (0)	1.8 (1.5-2.2, 0.8-5.8)
Absolute Monocytes *(10^9/L)*	0 (0)	0.3 (0.3-0.4, 0.1-1.3)	0 (0)	0.3 (0.3-0.4, 0.1-0.8)	0 (0)	0.4 (0.3-0.5, 0-1.3)	0 (0)	0.4 (0.3-0.5, 0-1.1)
Absolute Neutrophils *(10^9/L)*	0 (0)	4 (3.2-5.1, 0.9-11.6)	0 (0)	3.8 (3-4.7, 1.5-11.1)	0 (0)	4.1 (3.4-5.5, 1-15.2)	0 (0)	4.2 (3.5-5.2, 1.6-11.9)
Mean Cell Hemoglobin *(fmol)*	3 (1)	0.5 (0.5-0.5, 0.3-0.6)	4 (1)	0.5 (0.5-0.5, 0.3-0.6)	0 (0)	0.5 (0.4-0.5, 0.3-0.6)	0 (0)	0.5 (0.4-0.5, 0.3-0.6)
Mean Cell Volume *(fL)*	3 (1)	92 (89-96, 69-108)	4 (1)	92 (89-95, 70-113)	0 (0)	93 (90-97, 71-112)	0 (0)	93 (89-96, 65-110)
White Blood Cell *(total, 10^9/L)*	0 (0)	6.4 (5.5-7.7, 3-14.6)	0 (0)	6.3 (5.2-7.5, 3-14.2)	0 (0)	6.7 (5.6-8.2, 3.6-17.9)	0 (0)	6.9 (5.8-8, 3.3-14.8)
				Biochemistry				
Albumin *(g/L)*	21 (5)	46 (44-48, 38-53)	24 (6)	46 (44-48, 38-55)	0 (0)	46 (44-47, 37-61)	0 (0)	46 (43-47, 31-55)
Alkaline phosphatase *(serum, U/L)*	0 (0)	62 (49-75, 23-141)	0 (0)	61 (51-74, 28-141)	0 (0)	66 (54-81, 29-161)	0 (0)	71 (57-86, 1-159)
Creatinine *(µmol/L)*	0 (0)	68 (61-75.8, 38-124)	0 (0)	68 (60-78, 35-117)	0 (0)	67.2 (61-75.1, 41.5-114.9)	0 (0)	69 (61.2-79, 33.6-114.9)

	Development				External Validation			
	Placebo (*n* = 418)		Fingolimod (*n* = 425)		Placebo (*n* = 355)		Fingolimod (*n* = 358)	
Characteristic	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**	**Missing *n* (%)**	**Median (IQR, range), or *n* (%)**
Bilirubin *(direct/conjugated, µmol/L)*	0 (0)	2 (2-3, 0-10)	0 (0)	2 (2-3, 0-7)	0 (0)	1.7 (1.7-1.7, 0-5.1)	0 (0)	1.7 (1.7-1.7, 0-6.8)
Gamma Glutamyltransferase *(GGT, U/L)*	0 (0)	15 (11-21, 5-197)	0 (0)	15 (11-23, 4-193)	0 (0)	16 (12-24.5, 5-179)	0 (0)	17 (12-25, 6-180)
SGOT *(AST, U/L)*	0 (0)	19 (16-22, 10-92)	0 (0)	18 (16-22, 10-75)	0 (0)	18 (16-21, 9-75)	0 (0)	18 (16-23, 8-58)
SGPT *(ALT, U/L)*	0 (0)	17 (13-24, 6-146)	0 (0)	16 (13-23, 5-89)	0 (0)	18 (14-24, 5-151)	0 (0)	18 (14-25, 7-109)
Bilirubin *(total, µmol/L)*	0 (0)	8 (6-11, 2-50)	0 (0)	8 (6-11, 2-34)	0 (0)	6.8 (5.1-8.6, 1.7-22)	0 (0)	6.8 (5.1-8.6, 1.7-34.2)

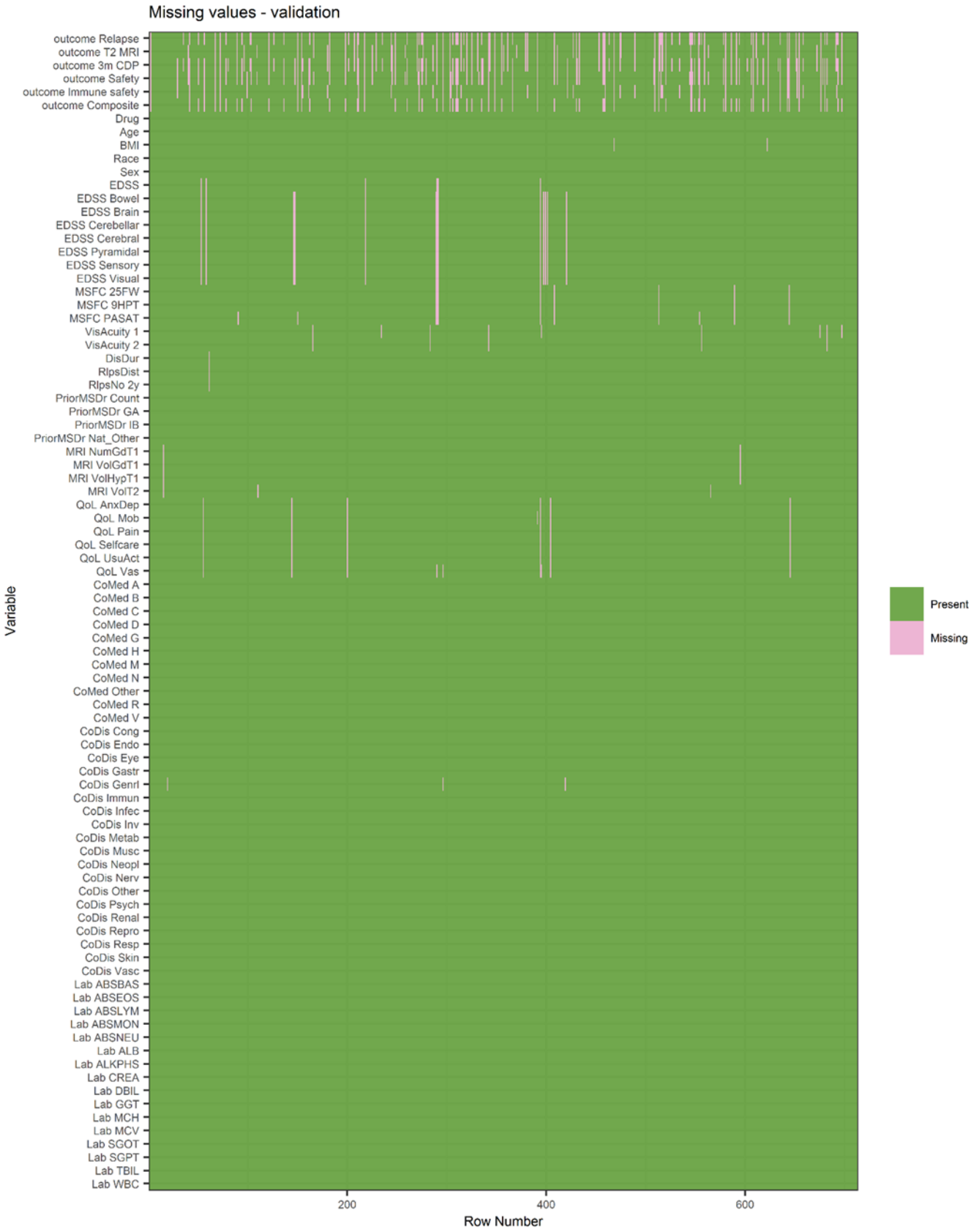
Table 5 *Baseline characteristics*

Description of all predictors in model development and external validation datasets by arm. Median (interquartile range, range) for numerical variables, and frequencies (percentage) for categorical variables and missing values.

Figure 6 a/b *Missing values*

Missing value pattern of predictors, and outcomes conceptualized as binary at 24 months. The rows, or participants, are represented in the x-axis. **a)** *This page* Model development dataset **b)** *Next page* External validation dataset.

.


Missing values - validation
outcome Relapse
outcome T2 MRI
outcome 3m CDP
outcome Safety
outcome Immune safety
outcome Composite
Drug
Age
BMI
Race
Sex
EDSS
EDSS Bowel
EDSS Brain
EDSS Cerebellar
EDSS Cerebral
EDSS Pyramidal
EDSS Sensory
EDSS Visual
MSFC 25FW
MSFC 9HPT
MSFC PASAT
VisAcuity 1
VisAcuity 2
DisDur
RIpsDist
RIpsNo 2y
PriorMSDr Count
PriorMSDr GA
PriorMSDr IB
PriorMSDr Nat_Other
MRI NumGdT1
MRI VolGdT1
MRI VolHypT1
MRI VolT2
QoL AnxDep
QoL Mob
QoL Pain
QoL Selfcare
QoL UsuAct
QoL Vas
CoMed A
CoMed B
CoMed C
CoMed D
CoMed G
CoMed H
CoMed M
CoMed N
CoMed Other
CoMed R
CoMed V
CoDis Cong
CoDis Endo
CoDis Eye
CoDis Gastr
CoDis Genrl
CoDis Immun
CoDis Infec
CoDis Inv
CoDis Metab
CoDis Musc
CoDis Neopl
CoDis Nerv
CoDis Other
CoDis Psych
CoDis Renal
CoDis Repro
CoDis Resp
CoDis Skin
CoDis Vasc
Lab ABSBAS
Lab ABSEOS
Lab ABSLYM
Lab ABSMON
Lab ABSNEU
Lab ALB
Lab ALKPHS
Lab CREA
Lab DBIL
Lab GGT
Lab MCH
Lab MCV
Lab SGOT
Lab SGPT
Lab TBIL
Lab WBC
Variable
200
400
600
Row Number
Present
Missing

.

4.2 Model development

Based on the cross-validated average cumulative AUC(t) (**Table 6**), the modeling methods with best discriminative power for predicting the relapse outcome were elastic net or grouped lasso and they had moderate performance ($AUC(t)_{avg}$ 0.69). Transformation forest did perform comparably ($AUC(t)_{avg}$ 0.64) but the transformation tree performed poorly ($AUC(t)_{avg}$ 0.50). The elastic net model was chosen as the preferred method predicting relapse due to the fact that it was more parsimonious with only five terms (**Table 7**) compared to the grouped lasso with 65 terms in the final model fits. It should be noted that the optimized alpha value in the elastic net was 1, indicating that only lasso penalization was used in the final model (**Table 8**). The predictors in the final model were all main terms: total EDSS score, total volume of Gd-enhanced T1 lesions, number of relapses in the last 2 years, and number of prior MS treatments, alongside the drug as fingolimod or placebo. The lack of treatment interaction terms in the best performing model led to question the extent to which time-to relapse was heterogeneous in response to fingolimod 0.5 mg, particularly at the relative scale.

The cross-validated discrimination performance of the transformation tree varied very close to 0.50 for all investigated outcomes in this study, indicating that its performance was almost equivalent to random choice. Both the elastic net and grouped lasso methods performed the best also for predicting new or enlarging T2 MRI lesions ($AUC(t)_{avg}$ 0.71) and, although with poorer discrimination, immunosuppressant safety ($AUC(t)_{avg}$ 0.60). Elastic net was the simpler model for these outcomes (9 and 45 terms compared to 19 and 81 terms of grouped lasso), too, so it was chosen as the final model over grouped lasso or transformation forest that had performance close to the penalized regression algorithms. The coefficients and the baseline cumulative hazard for the final model fits of the chosen regression methods can be found in Appendix B. Inference about treatment effect neither was an objective of this study nor is appropriate with the methods used. Still, the coefficients of treatment revealed that it decreased the risk of the event with respect to all the four outcomes for which the final model was a regression, especially for relapse and new/enlarging lesions.

In predicting time-to 3-month CDP, the transformation forest greatly outperformed the other methods ($AUC(t)_{avg}$ 0.67) with more than 0.1 difference in discriminative performance. Transformation forest had a very low discrimination ability ($AUC(t)_{avg}$ 0.54) for the safety outcome but it still was the best among the others. Because the safety and efficacy outcome was a composite of others for which the models had moderate or poor discriminative power, the performance of the methods for the composite outcome

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety	Composite
Transformation tree	0.50	0.47	0.54	0.51	0.54	0.49
Transformation forest	0.64	0.68	0.67	0.54	0.60	0.59
Elastic net	0.69	0.71	0.56	0.51	0.60	0.61
Grouped lasso	0.69	0.71	0.55	0.50	0.60	0.63

Table 6 *Cross-validated area under the curve*

Average cumulative time-dependent area under the curve at 6, 12, and 24 months estimated via cross-validation in the model development dataset. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety	Composite
Transformation tree	3	3	0	2	2	2
Elastic net	5	9	11	2	45	17
Grouped lasso	65	19	35	17	81	25

Table 7 *Number of predictors in competing models*
The number of splits (transformation tree) or terms (elastic net and grouped lasso) in the model fits. The number of splits in transformation trees is not directly comparable to the number of terms chosen in penalized regression methods because the tree has an internal interaction structure. The random forest algorithm does not have variable selection, so the transformation forests had all 80 predictors in interaction with treatment for all the outcomes. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

was somewhere in between. The grouped lasso, containing 25 terms, was the modeling method with the highest $AUC(t)_{avg}$ of 0.63 for predicting the composite outcome.

The cross-validated AUC(t) at months 6, 12, and 24 are provided separately in Appendix B. The cross-validated average $Brier(t)_{avg}$ demonstrated that the models with the lowest $Brier(t)_{avg}$ scores were the ones selected by using $AUC(t)_{avg}$ except for the safety model, $Brier(t)_{avg}$ score of which was only 0.001 lower with elastic net compared to with transformation forest (**Table 9**). Hence, the results from the model development stage can be considered robust to the performance evaluation method.

4.3 Variable importance

For predicting relapse, the predictor total volume of Gd-enhanced T1 lesions was found to be important by all modeling methods during cross-validation and it was also selected by lasso penalty in the final model fit. The predictors that were deemed important by three of the four modeling methods during cross-validation were total volume of T2 lesions and the concomitant diseases from the SOC of metabolism and nutrition disorders. Yet, these predictors were not selected in the final model fit. The other predictors selected in the final model (total EDSS score, number of relapses in the last 2 years, and number of prior MS treatments) were chosen at least twice by both of the penalized regression methods during cross-validation.

Outcome	Method	Parameters
Relapse	elastic net	alpha=1, lambda=0.06
New/enlarging lesions	elastic net	alpha=0.1, lambda=0.59
Confirmed disability progression	transformation forest	mtry=8, minbucket=10
Safety	transformation forest	mtry=27, minbucket=15
Immunosuppressant safety	elastic net	alpha=0.1, lambda=0.32
Safety and efficacy	grouped lasso	alpha=0.6, lambda=0.07

Table 8 *Final methods and tuning parameters*
The final modeling methods chosen for each outcome and the tuning parameters of the final model.

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety	Composite
Transformation tree	0.224	0.292	0.118	0.086	0.220	0.243
Transformation forest	0.199	0.208	0.115	0.085	0.215	0.231
Elastic net	0.181	0.197	0.118	0.084	0.215	0.219
Grouped lasso	0.182	0.210	0.122	0.085	0.240	0.218

Table 9 *Cross-validated Brier score*
Average time dependent Brier score at 6, 12, and 24 months estimated via cross-validation in the development dataset. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

The top three important predictors for predicting new or enlarging T2 MRI lesions were other MRI parameters. Total volume and number of Gd-enhanced T1 lesions were important irrespective of the modeling method and total volume of T2 lesions was found to be important by three of the four methods. These three predictors were also selected in the final model fit. Other predictors in the final model fit (age, duration of MS since 1st symptom, QoL: visual analog scale, and bilirubin) were chosen at least twice by both of the penalized regression methods during cross-validation.

For predicting CDP, no splitting variable was selected by the transformation tree at least twice out of the five cross-validation folds, hinting that the structure of a tree may be too simplistic for modeling this complex outcome. The remaining three methods chose mean of 9HPT from the MSFC panel and the concomitant diseases from the SOC of musculoskeletal and connective tissue disorders as important to predict CDP. The single predictor deemed important by all methods for predicting the safety outcome was the concomitant disease of gastrointestinal disorders. In predicting immunosuppressant safety, no variable was found important by the transformation forest at least twice out of the five cross-validation folds. The remaining three methods found exposure to comedications of genito urinary system and sex hormones as important to predict the risk of infections or neoplasms. Finally, the single predictor that was deemed important to predict the composite safety and efficacy outcome by all methods was total volume of Gd-enhanced T1 lesions. **Figure 7** and **Figure 8** demonstrate how important variables were deduced for the transformation tree and forest algorithms. The list of all predictors deemed important by modeling method per each outcome are provided in Appendix B.

4.4 External validation

The final models were used to predict the probability of the outcomes in the external validation dataset based on the baseline predictors and the actual treatment arm. These predictions (**Table 10** and **Figure 9**) and the observed outcomes were used to evaluate the discrimination and calibration in Section 4.4.1 and the net benefit of the models in Section 4.4.2. At month 24, median individual prediction for relapse risk was 0.42 (range 0.21-0.87). The highest predicted risk distribution was that of experiencing the immunosuppressant safety endpoint (median 0.86, range 0.59-1.00) whereas the lowest predicted risk distribution was that of experiencing the overall safety endpoint (median 0.16, range 0.06-0.34).

Finally, the counterfactual outcomes were predicted assuming the participants were assigned to the treatment arm other than theirs. These were used to evaluate the predicted treatment response in Section 4.4.2 which was calculated in an individual patient by predicting the risk of an outcome

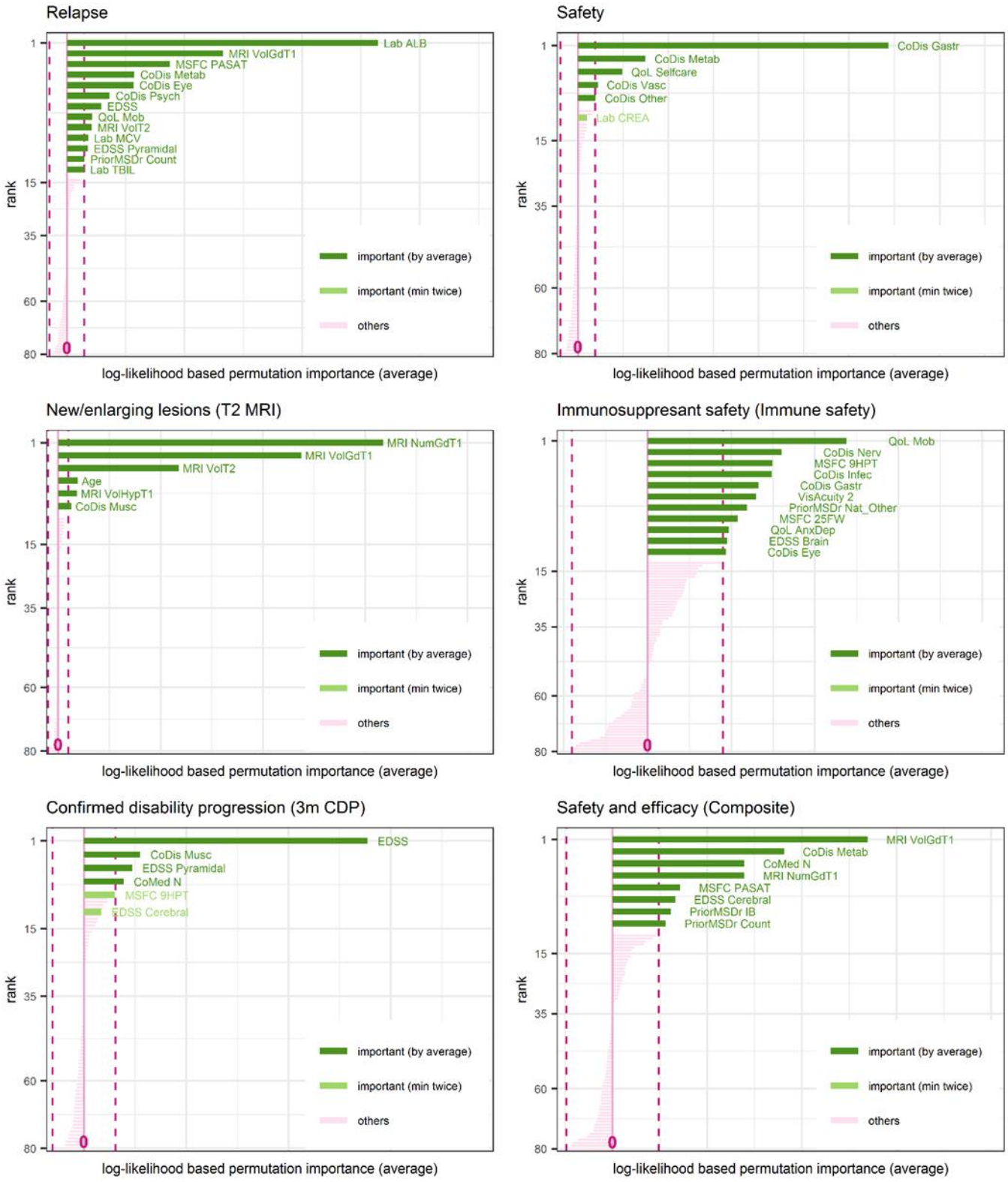


Figure 7 *Variable importance from transformation forests*

Plots of variable importance from the transformation forest algorithm developed in the training set and evaluated by the log-likelihood based permutation importance in the test set of the five folds of cross validation. Dark green indicates the important predictors when the log-likelihood averaged over the folds of is considered. The predictors which were not important based on the average but were important in at least two folds are light green. Other predictors are represented in pink. CoDis: concomitant disease, CoMed: comedication.

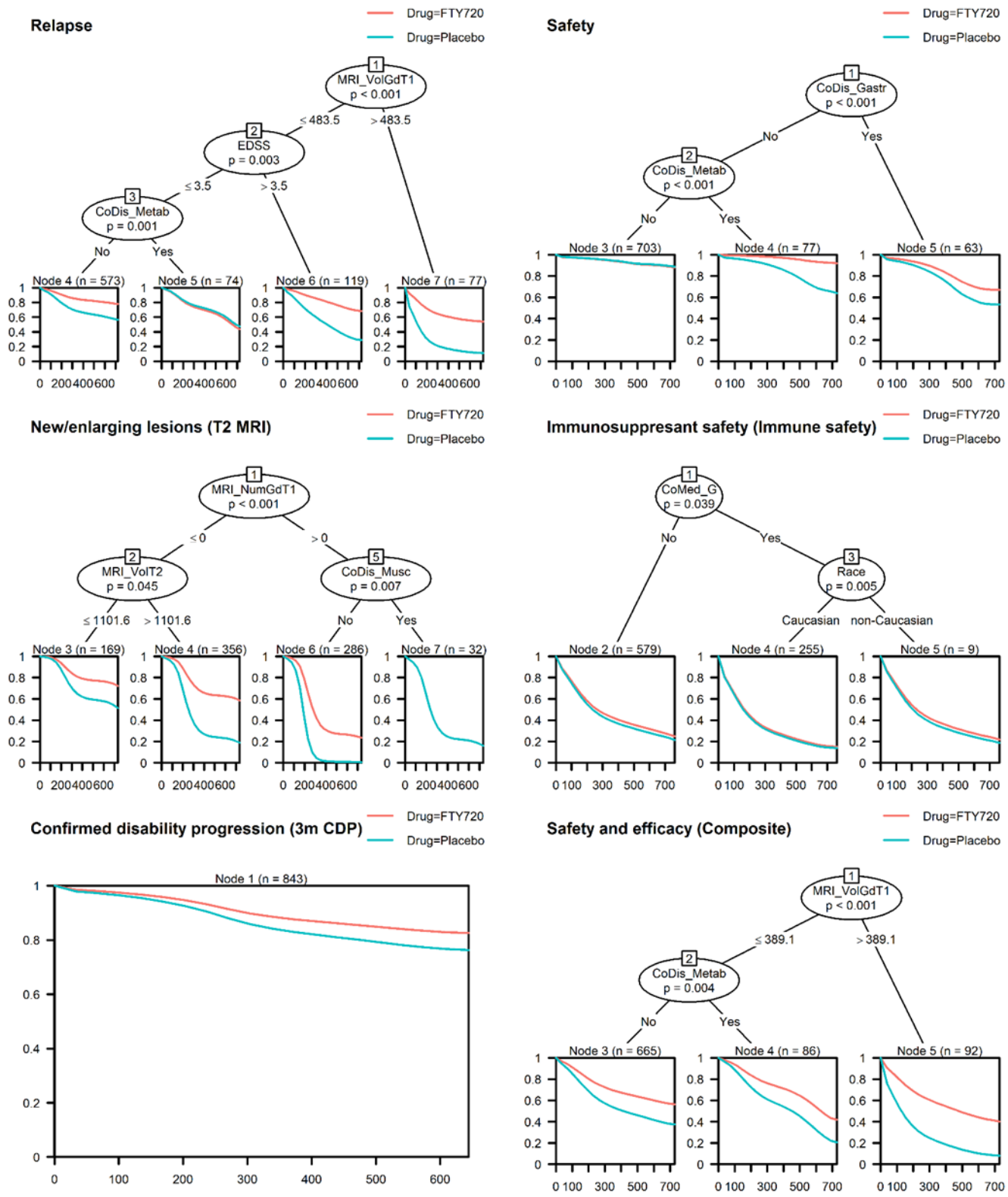


Figure 8 *Transformation trees*

Plots of transformation trees fit to all of the development dataset. None of these were chosen as the final model for the specific outcomes. The figures describe the survival probability in the subgroup belonging to the node per trial arm: active fingolimod 0.5 mg as FTY720 and control arm as Placebo. In some nodes (e.g. Node 5 of the tree for Relapse), the curves for the arms are indistinguishable when there is almost no difference in survival. MRI: Magnetic resonance imaging, CoDis: Concomitant disease, CoMed: Comedication.

Outcome	Overall Median (range)	FTY720 Median (range)	Placebo Median (range)
Relapse	0.42 (0.21-0.87)	0.28 (0.21-0.71)	0.53 (0.42-0.87)
New/enlarging lesions	0.68 (0.38-0.98)	0.47 (0.38- 0.87)	0.76 (0.69-0.98)
Confirmed disability progression	0.22 (0.10-0.37)	0.23 (0.10-0.34)	0.22 (0.11-0.37)
Safety	0.16 (0.06-0.34)	0.16 (0.06-0.33)	0.16 (0.07-0.34)
Immunosuppressant safety	0.86 (0.59-1.00)	0.86 (0.59-1.00)	0.86 (0.70-0.99)
Safety and efficacy	0.59 (0.39-1.00)	0.47 (0.39-0.69)	0.66 (0.58-1.00)

Table 10 *Predicted event probabilities*

Summary of predicted individual event probabilities at 24 months derived from the final models in the external validation dataset, overall and by treatment arms. FTY720: fingolimod.

separately under placebo and under fingolimod 0.5 mg and taking their difference. At 24 months, median predicted individual reduction in relapse risk by daily fingolimod 0.5 mg compared to placebo (**Table 11**) was 0.25 (range 0.21-0.31). The highest median predicted individual risk reduction by fingolimod was in the risk of new or enlarging T2 MRI lesions (median 0.29, range 0.12-0.32). According to the summary measures of predicted individual risk change, the median response to daily fingolimod 0.5 mg compared to placebo was null (minimum and maximum almost symmetric around null) for CDP, safety, and immunosuppressant safety outcomes.

4.4.1 Discrimination and calibration

The discriminative performance of the final model predicting relapse in the external validation dataset was very close to the cross-validated performance in the model development dataset. The cumulative AUC at 24 months was 0.68 (95% CI: 0.63-0.72). The discrimination performances at months 6 and 12 were similar (**Table 12**, **Figure 10**, and **Figure 11**). The improvement in Brier score (**Table 13**) was 7% at month 24 and the plots of monthly Brier score of the final relapse model compared to that of the null

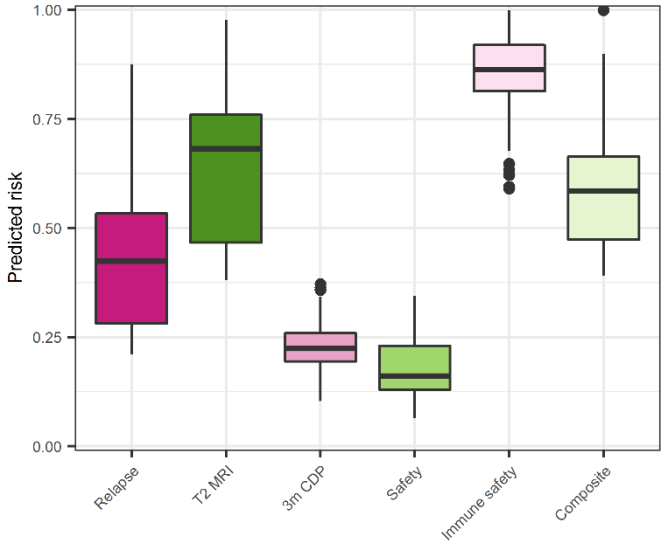


Figure 9 *Predicted event probabilities*

Boxplots of predicted individual event probabilities at 24 months derived from the final models in the external validation dataset, P(Outcome | Drug & all predictors).

Outcome	Median (range)
Relapse	0.25 (0.21-0.31)
New/enlarging lesions	0.29 (0.12-0.32)
Confirmed disability progression	0.00 (-0.18-0.18)
Safety	0.00 (-0.23-0.23)
Immunosuppressant safety	0.00 (-0.13-0.12)
Safety and efficacy	0.19 (0.10-0.45)

Table 11 *Predicted treatment response*

Summary of predicted individual response to fingolimod at 24 months derived from the final models in the external validation dataset. The predicted treatment response for an individual participant is calculated by predicting the risk of outcome when the drug is fingolimod 0.5 mg and taking its difference from the predicted risk of outcome when the drug is placebo. For example, for a patient with given baseline characteristics, the risk prediction for experiencing a relapse within the next 2 years can be 0.62 under placebo, but 0.38 under fingolimod. Then, the predicted treatment response of absolute relapse risk reduction for this patient would be 0.62-0.38=0.24.

model revealed an overall fit which may not be significantly different than that of the null model. The calibration plot (**Figure 11**) and calibration-in-the-large (**Table 14**) (-0.17, 95% CI -0.3 - -0.04)) revealed significant overestimation (observed/expected 0.84) of the relapse risk by the model in the external validation dataset. The bins in the calibration plots varied around the diagonal line and the calibration slope (1.06, 95% CI 0.78-1.35) was very close to 1 indicating that a change in the predicted risk may lead to a slightly higher change in the actual risk. Interestingly, the predicted risks of the binned groups per treatment arm were completely separate while their estimated actual risk had small overlap, revealing the great influence of the drug in the final model.

The external validation performance of the final model predicting new or enlarging lesions in T2 MRI (AUC at 24 months: 0.74, 95% CI 0.70-0.78) was also very close to the cross-validated performance in the model development dataset. The overall fit of the model for this outcome was not good but its discrimination performance was the best compared to the other outcomes. The improvement in Brier score was 12% at month 24 and the 95% confidence intervals of the final and null models for predicting new or enlarging T2 MRI lesions did not overlap after about 7 months. The calibration-in-the-large (-0.08, 95% CI -0.18-0.01) revealed that the model slightly but non-significantly overestimated (observed/expected 0.92) the probability of having new or enlarging lesions. Similar to that of relapse,

Outcome	Month 6	Month 12	Month 24
Relapse	0.65 (0.60-0.71)	0.68 (0.63-0.73)	0.68 (0.63-0.72)
New/enlarging lesions	0.63 (0.58-0.69)	0.73 (0.69-0.76)	0.74 (0.70-0.78)
Confirmed disability progression	0.74 (0.68-0.81)	0.65 (0.59-0.71)	0.59 (0.54-0.64)
Safety	0.45 (0.37-0.53)	0.45 (0.39-0.52)	0.50 (0.44-0.55)
Immunosuppressant safety	0.59 (0.54-0.63)	0.61 (0.56-0.66)	0.69 (0.63-0.74)
Safety and efficacy	0.58 (0.53-0.62)	0.59 (0.55-0.63)	0.58 (0.53-0.63)

Table 12 *Area under the curve*

Cumulative time-dependent area under the curve with uncertainty (95% confidence interval) at 6, 12, and 24 months estimated for the final models in the external validation dataset.

Outcome	Month 6	Month 12	Month 24
Relapse	4%	6%	7%
New/enlarging lesions	2%	10%	12%
Confirmed disability progression	1%	3%	1%
Safety	-3%	-3%	-3%
Immunosuppressant safety	0%	1%	5%
Safety and efficacy	0%	2%	0%

Table 13 *Scaled Brier score*

Scaled time-dependent Brier scores at 6, 12, and 24 months in the external validation dataset showing percentage improvement in the Brier score by the final models compared to that of the null model.

the confidence interval of the calibration slope (1.07, 95% CI 0.83-1.31) contained 1 indicating that a change in the predicted risk may lead to only a slightly higher change in the actual risk. Also similar to the relapse outcome, the predicted risk of the groups per treatment arm were completely separate although their actual risk overlapped, revealing how influential treatment was in the developed prediction model. Another interesting observation from the calibration plot was the very small range, about 15%, of predicted risk within treatment arms, although the actual risk had a range of about 50%. The observably large slope when stratified by treatment arms might indicate that conditional on treatment, one unit change in the predicted risk corresponded to a change that was greater than one unit in the actual risk.

The transformation forest predicting CDP had a similar discrimination performance in the external validation compared to that in the development dataset at 12 months (AUC: 0.65, 95% CI 0.59-0.71). However, more pronounced was its increased performance to predict timespans closer to (AUC at 6 months: 0.74, 95% CI 0.68-0.81) and decreased performance to predict timespans farther away than (AUC at 24 months: 0.59, 95% CI 0.54-0.64) the baseline. The Brier score of the final model was only marginally better than the null model (scaled Brier score 1%) and their monthly plots were almost overlapping. According to the calibration-in-the-large (0.17, 95% CI 0.02-0.32), significantly more (observed/expected 1.19), patients than predicted by the model experienced CDP in the external validation dataset. Unlike the calibration plots of relapse or new or enlarging lesions in T2 MRI, the predictions by treatment arms were not clearly separated.

Outcome	Calibration-in-the-large	Calibration slope
Relapse	-0.17 (-0.3 - -0.04)	1.06 (0.78-1.35)
New/enlarging lesions	-0.08 (-0.18 - 0.01)	1.07 (0.83-1.31)
Confirmed disability progression	0.17 (0.02 - 0.32)	-
Safety	0.07 (-0.11 - 0.24)	-
Immunosuppressant safety	-0.15 (-0.24 - -0.07)	0.66 (0.43-0.89)
Safety and efficacy	0.07 (-0.03 - 0.16)	0.46 (0.18-0.74)

Table 14 *Calibration measures*

Calibration in-the-large and slopes of the final models in the external validation dataset, with uncertainty (95% confidence interval). Calculations for the slope are based on Cox models and estimated only for those methods with linear predictors.

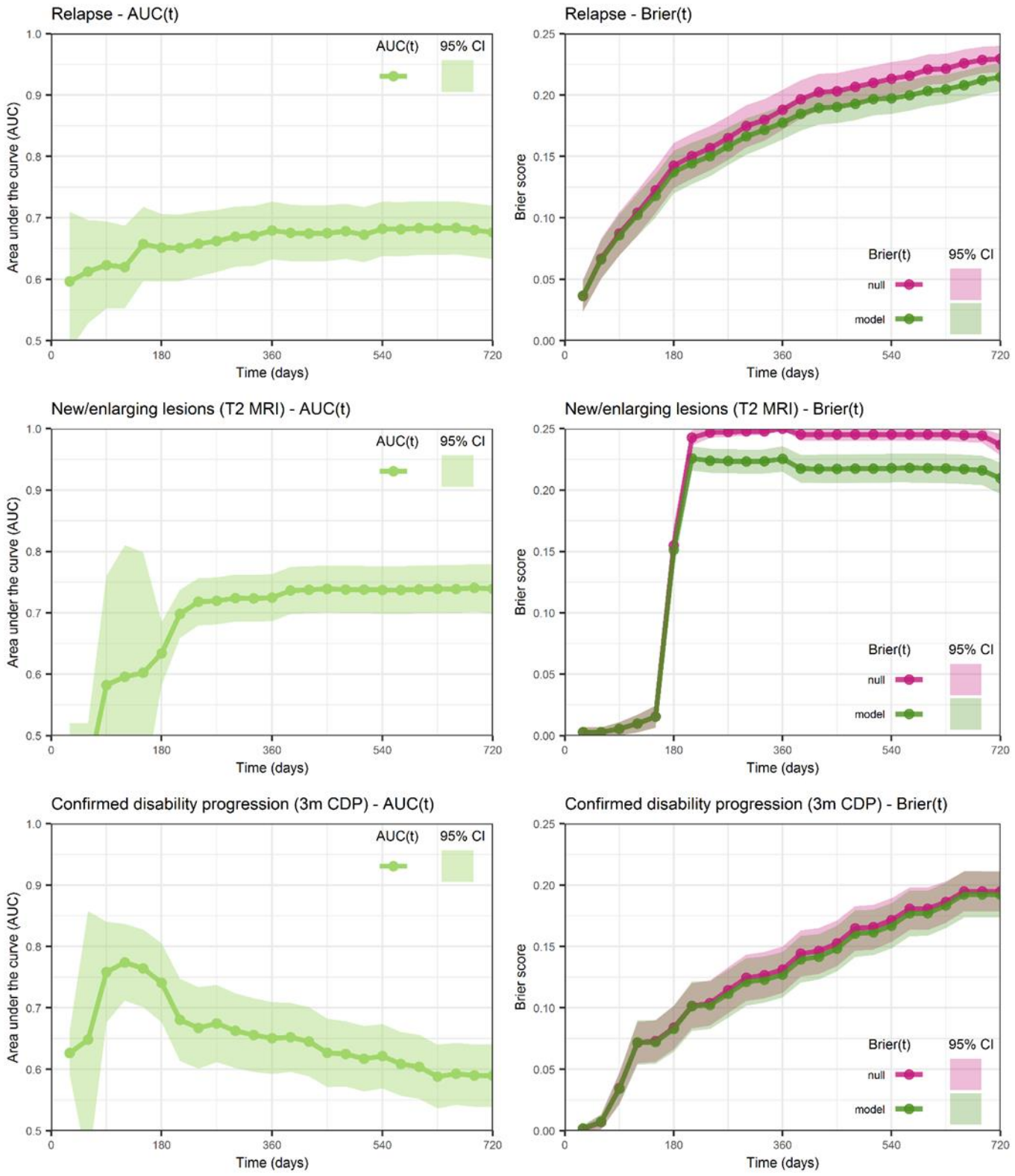
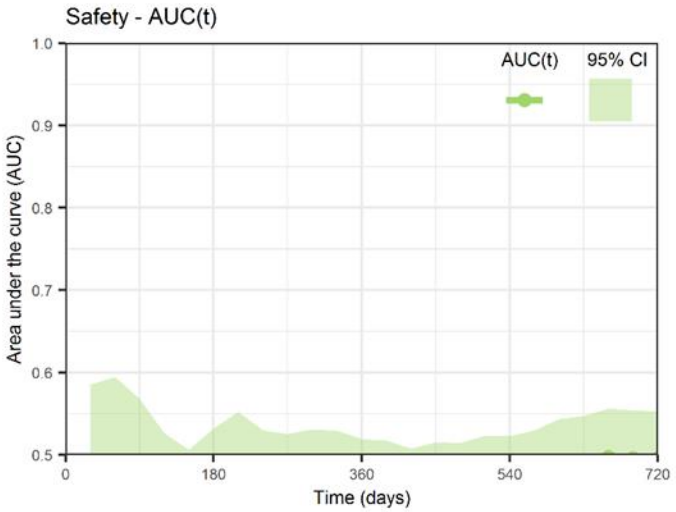
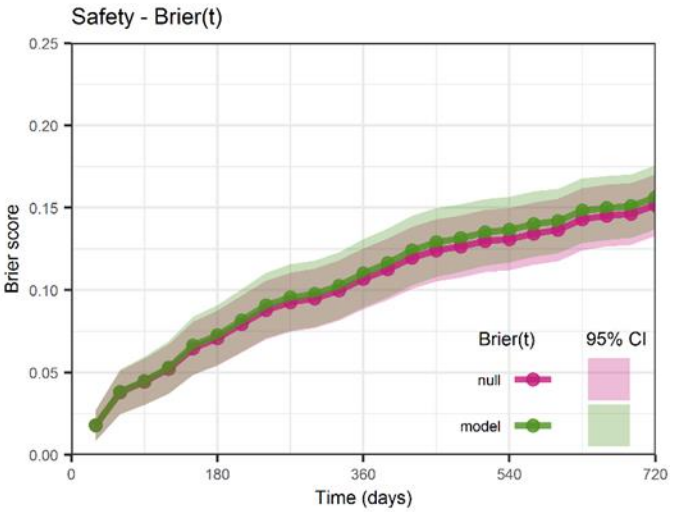
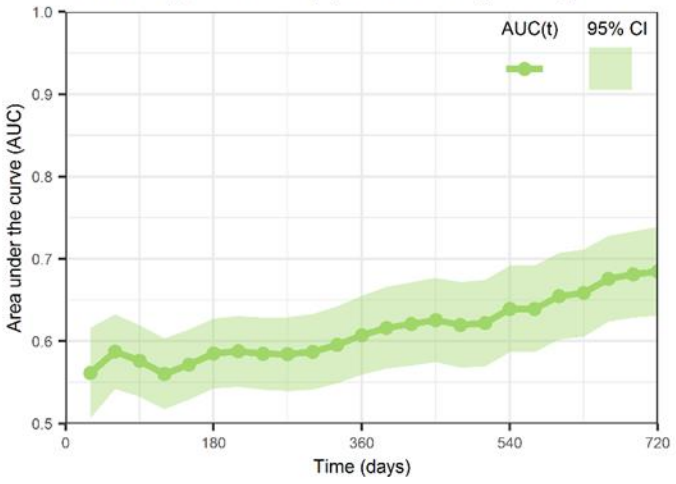


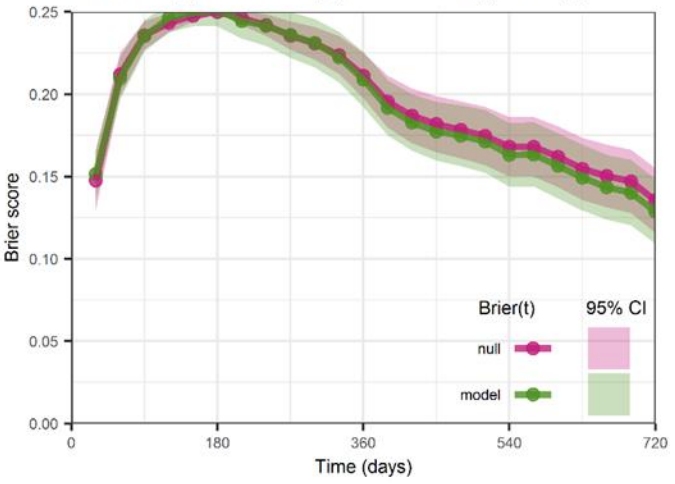
Figure 10 a/b *Area under the curve and Brier score over time*

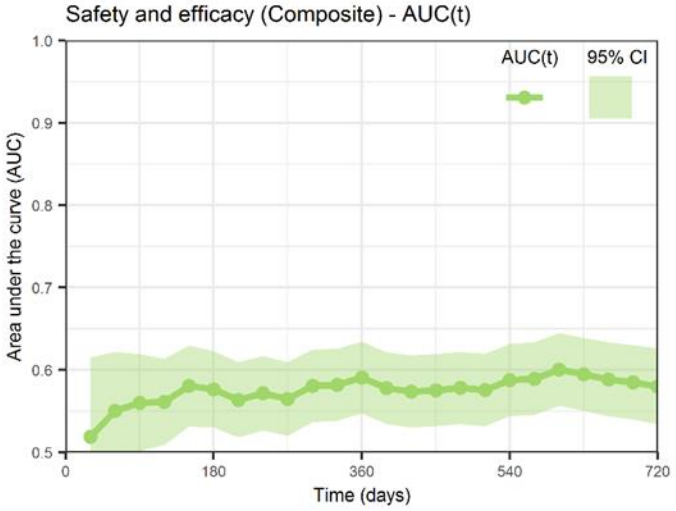
Monthly time-dependent cumulative area under the curve (AUC(t)) and Brier(t) scores with uncertainty (95% confidence intervals) of the final models for all outcomes in the external validation dataset. The plots of Brier(t) contain the null models (null) as a reference to the final models (model). **a)** *This page*: Relapse, New/enlarging lesions (T2 MRI), and Confirmed disability progression (3m CDP) **b)** *Next page*: Safety, Immunosuppressant safety (Immune safety), and Safety and efficacy (Composite).

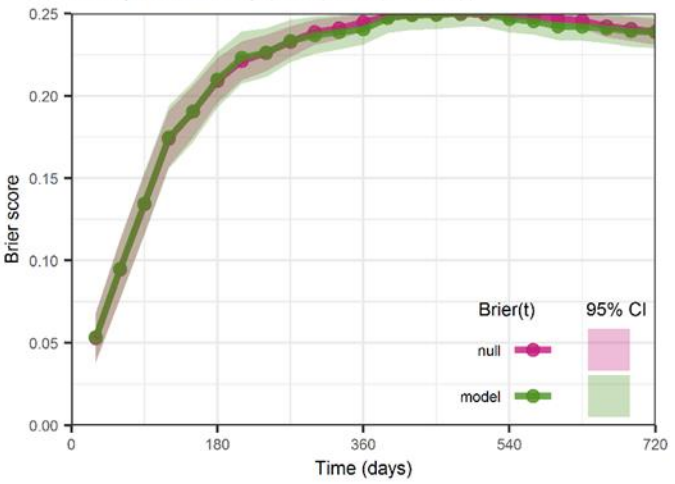

Safety - AUC(t)
AUC(t)
95% CI
Area under the curve (AUC)
1.0
0.9
0.8
0.7
0.6
0.5
0
180
360
540
720
Time (days)


Safety - Brier(t)
Brier score
0.25
0.20
0.15
0.10
0.05
0.00
Brier(t)
95% CI
null
model
0
180
360
540
720
Time (days)


Immunosuppresant safety (Immune safety) - AUC(t)
AUC(t)
95% CI
Area under the curve (AUC)
1.0
0.9
0.8
0.7
0.6
0.5
0
180
360
540
720
Time (days)


Immunosuppresant safety (Immune safety) - Brier(t)
Brier score
0.25
0.20
0.15
0.10
0.05
0.00
Brier(t)
95% CI
null
model
0
180
360
540
720
Time (days)


Safety and efficacy (Composite) - AUC(t)
AUC(t)
95% CI
Area under the curve (AUC)
1.0
0.9
0.8
0.7
0.6
0.5
0
180
360
540
720
Time (days)


Safety and efficacy (Composite) - Brier(t)
Brier score
0.25
0.20
0.15
0.10
0.05
0.00
Brier(t)
95% CI
null
model
0
180
360
540
720
Time (days)

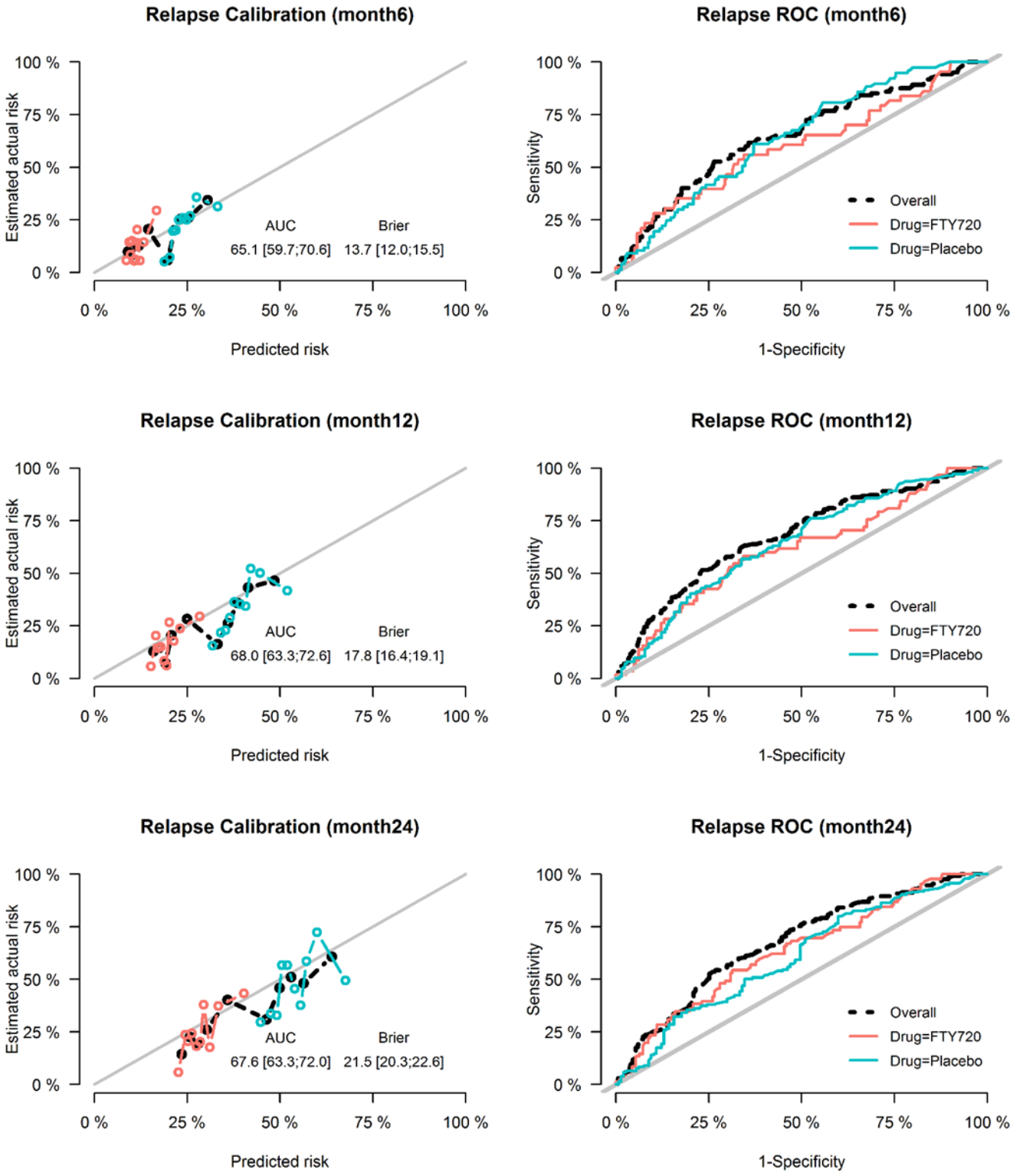
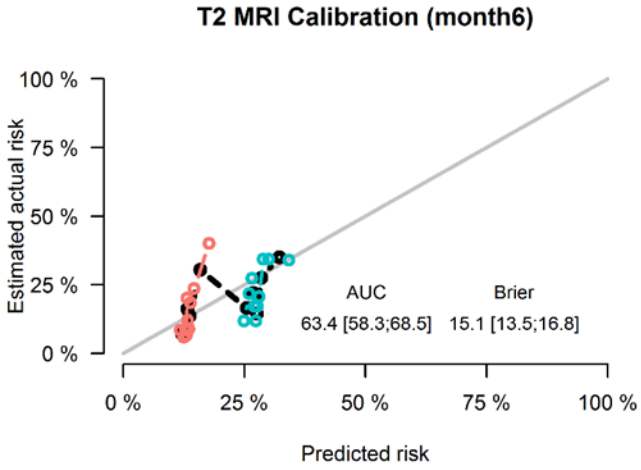
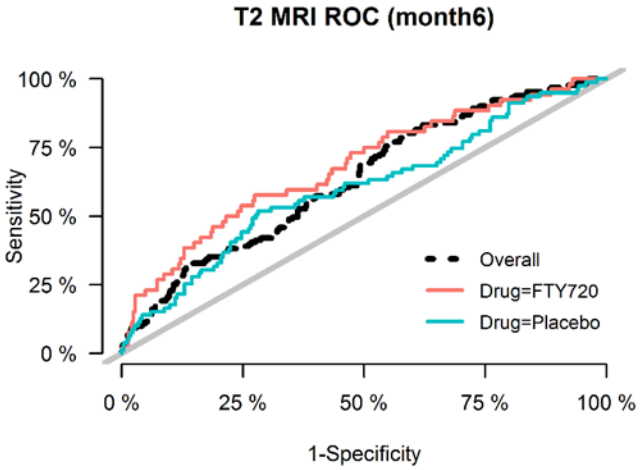
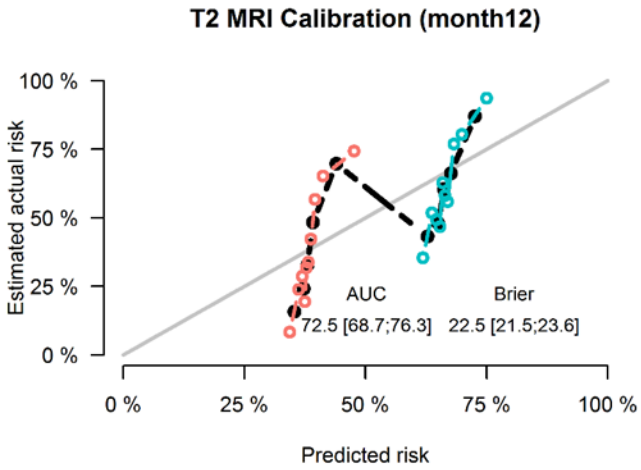


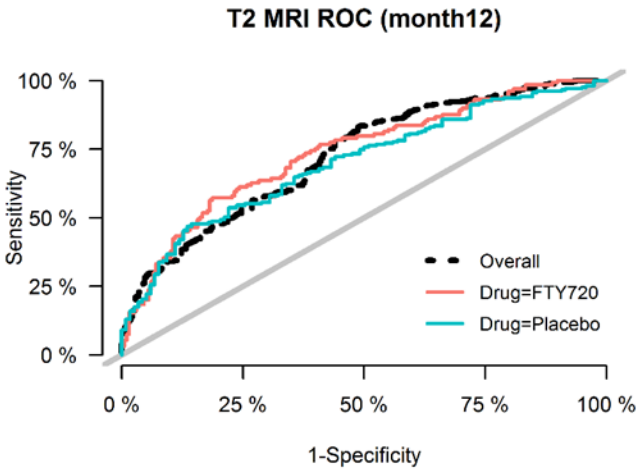
Figure 11 a/b *Calibration and receiver operator characteristic plots*

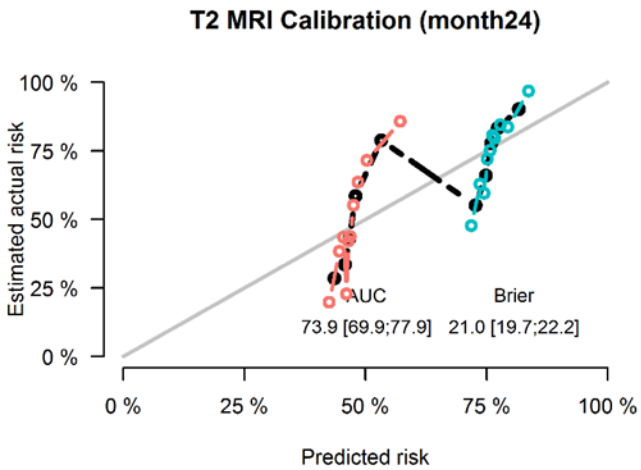
Calibration (binned to ten predicted risk groups) plots and receiver operator characteristic (ROC) curves overall and stratified by trial arm (active fingolimod 0.5 mg arm as FTY720 and control arm as Placebo) at months 6, 12, and 24 in the external validation dataset. Also provided are time-dependent area under the curve (AUC(t)) and Brier score (Brier(t)) as % at the respective time points. **a)** *This page* Relapse; **b)** *Next page*. New/enlarging lesions (T2 MRI).

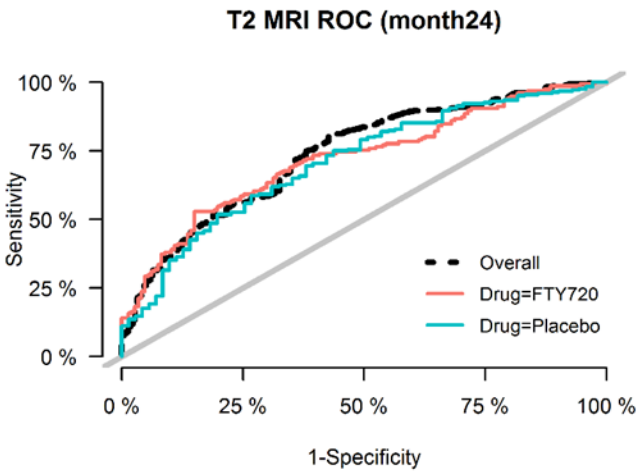

T2 MRI Calibration (month6)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
63.4 [58.3;68.5]
15.1 [13.5;16.8]
0 %
25 %
50 %
75 %
100 %
Predicted risk


T2 MRI ROC (month6)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


T2 MRI Calibration (month12)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
72.5 [68.7;76.3]
22.5 [21.5;23.6]
0 %
25 %
50 %
75 %
100 %
Predicted risk


T2 MRI ROC (month12)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


T2 MRI Calibration (month24)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
73.9 [69.9;77.9]
21.0 [19.7;22.2]
0 %
25 %
50 %
75 %
100 %
Predicted risk


T2 MRI ROC (month24)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity

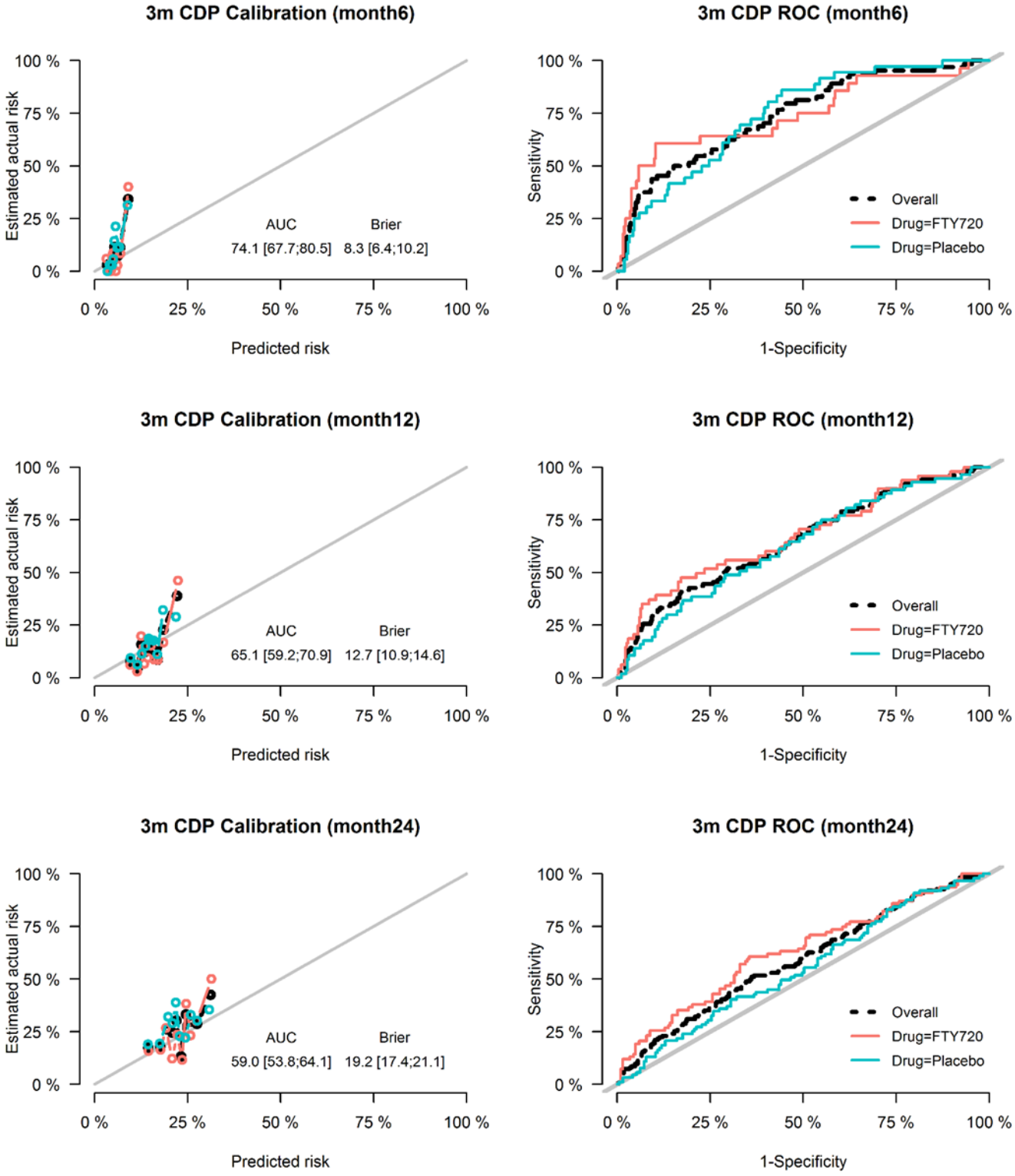
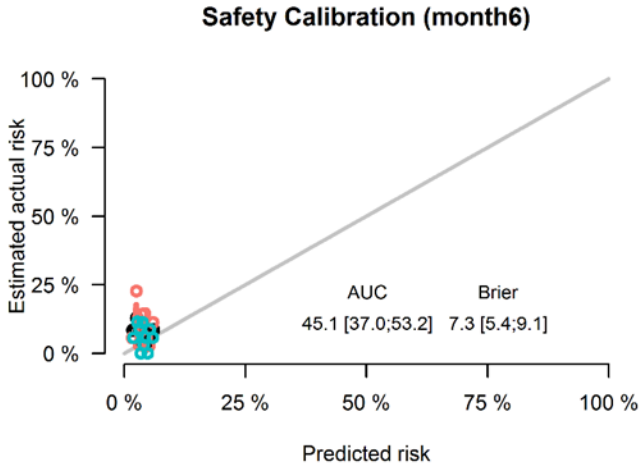
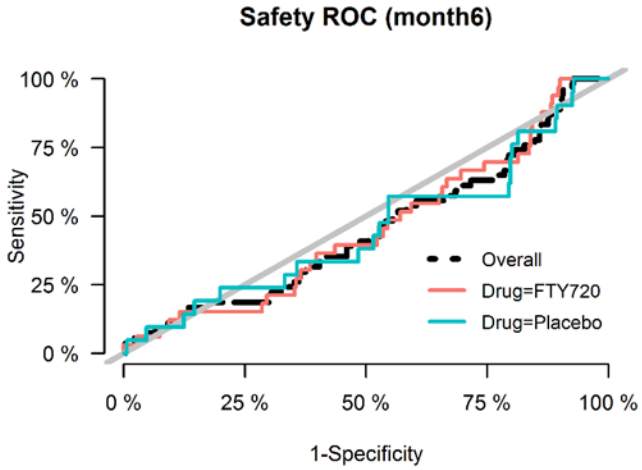
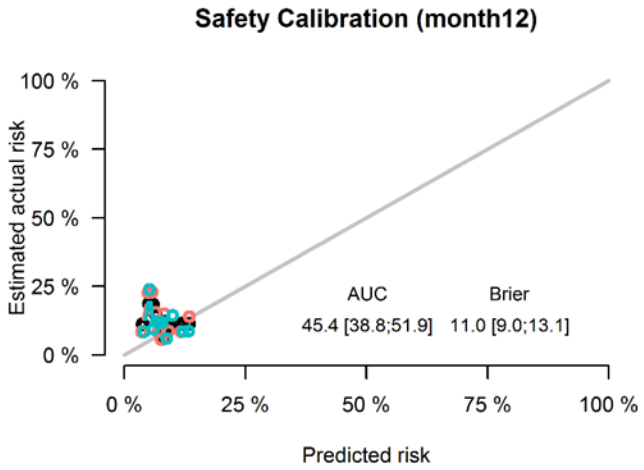


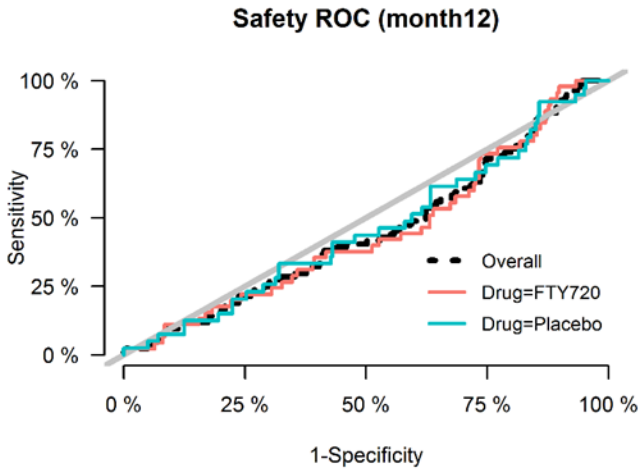
Figure 11 c/d *Calibration and receiver operator characteristic plots*

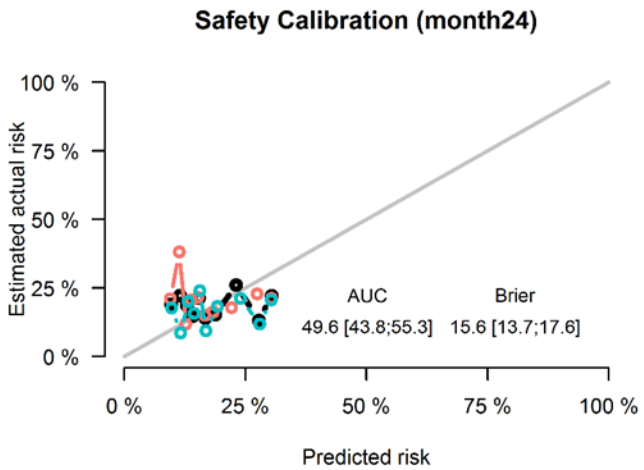
Calibration (binned to ten predicted risk groups) plots and receiver operator characteristic (ROC) curves overall and stratified by trial arm (active fingolimod 0.5 mg arm as FTY720 and control arm as Placebo) at months 6, 12, and 24 in the external validation dataset. Also provided are time-dependent area under the curve (AUC(t)) and Brier score (Brier(t)) as % at the respective time points. **c)** *This page* Confirmed disability progression (3m CDP); **d)** *Next page* Safety.

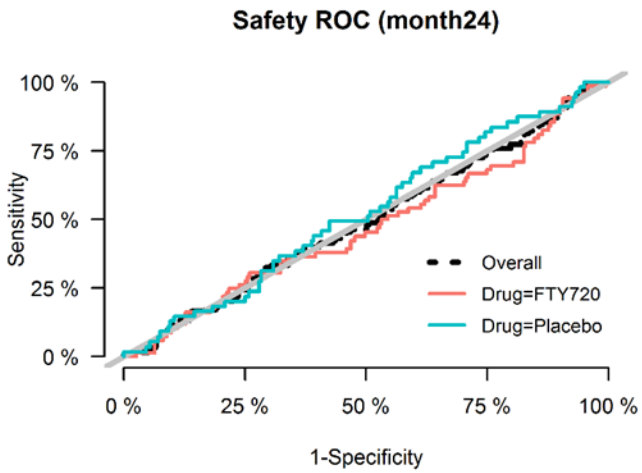

Safety Calibration (month6)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
45.1 [37.0;53.2]
7.3 [5.4;9.1]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Safety ROC (month6)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


Safety Calibration (month12)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
45.4 [38.8;51.9]
11.0 [9.0;13.1]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Safety ROC (month12)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


Safety Calibration (month24)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
49.6 [43.8;55.3]
15.6 [13.7;17.6]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Safety ROC (month24)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity

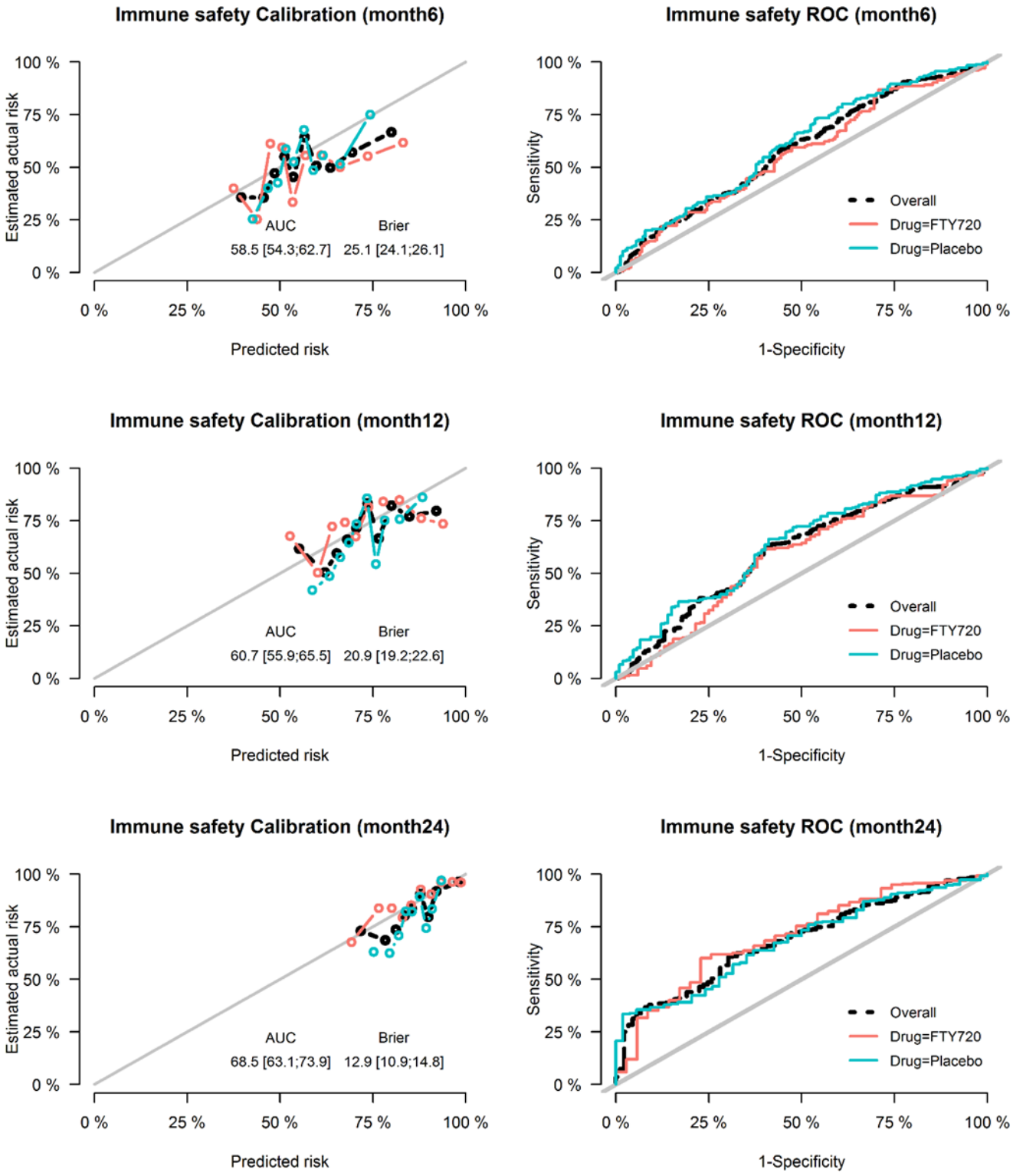
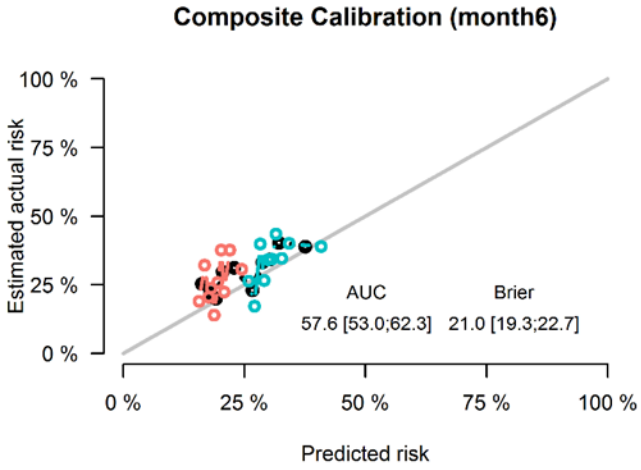
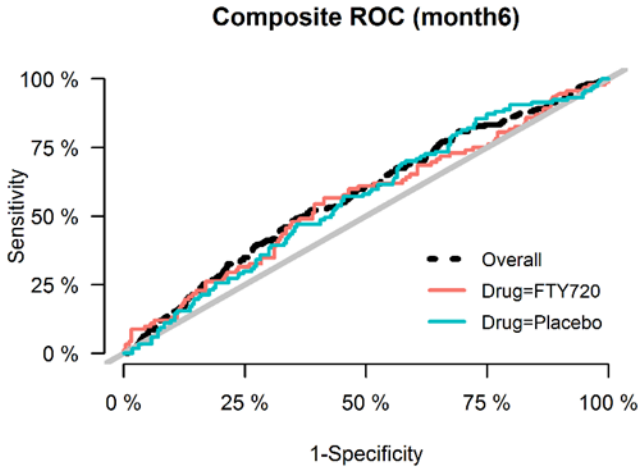
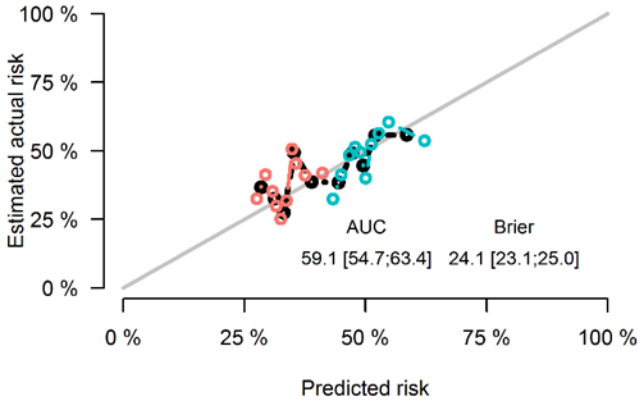


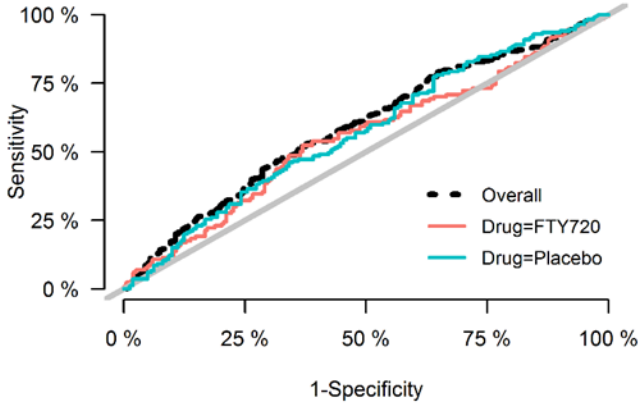
Figure 11 e/f *Calibration and receiver operator characteristic plots*

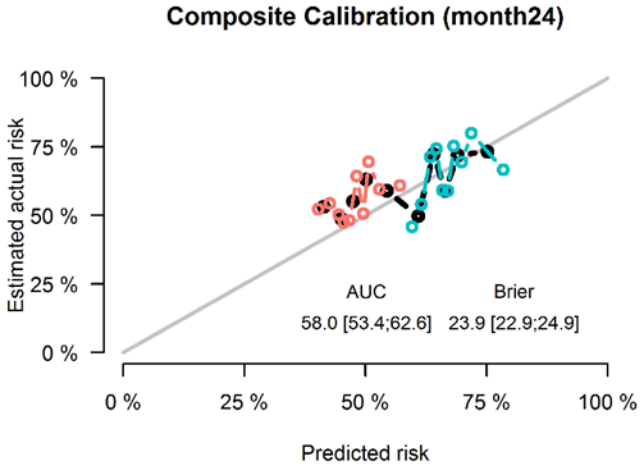
Calibration (binned to ten predicted risk groups) plots and receiver operator characteristic (ROC) curves overall and stratified by trial arm (active fingolimod 0.5 mg arm as FTY720 and control arm as Placebo) at months 6, 12, and 24 in the external validation dataset. Also provided are time-dependent area under the curve (AUC(t)) and Brier score (Brier(t)) as % at the respective time points. **e)** *This page* Immunosuppressant safety (Immune safety); **f)** *Next page* Safety and efficacy (Composite).

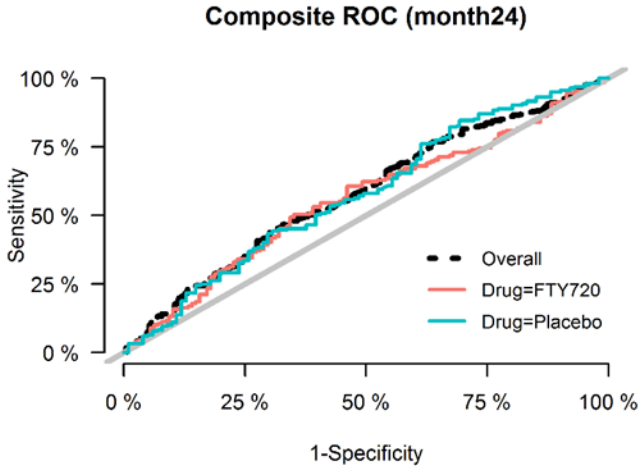

Composite Calibration (month6)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
57.6 [53.0;62.3]
21.0 [19.3;22.7]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Composite ROC (month6)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


Composite Calibration (month12)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
59.1 [54.7;63.4]
24.1 [23.1;25.0]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Composite ROC (month12)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity


Composite Calibration (month24)
Estimated actual risk
100 %
75 %
50 %
25 %
0 %
AUC
Brier
58.0 [53.4;62.6]
23.9 [22.9;24.9]
0 %
25 %
50 %
75 %
100 %
Predicted risk


Composite ROC (month24)
Sensitivity
100 %
75 %
50 %
25 %
0 %
Overall
Drug=FTY720
Drug=Placebo
0 %
25 %
50 %
75 %
100 %
1-Specificity

The final model for the safety outcome did not have any discriminative power in the external validation dataset: the AUC(t) for all time points contained 0.5 in their confidence interval (AUC at 24 months: 0.50, 95% CI 0.44-0.55). Also, the Brier score of the final model was worse than the null model (scaled Brier score at 24 months: -3%). Although the uncertainty around calibration-in-the-large of the safety outcome contained 0 (0.07, 95% CI -0.11-0.24), the calibration plots showed a very narrow range of estimated actual risk and no observable correlation with predicted risk.

The discriminative power of the immunosuppressant safety at 24 months was slightly better in the external validation dataset (AUC at 24 months: 0.69, 95% CI 0.63-0.74) compared to the cross-validated performance from the development dataset. Yet, the improvement in Brier score at month 24 was poor (5%) and the Brier score of the final model was not significantly different than that of the null. According to the calibration plot and the calibration-in-the-large (-0.15, 95% CI -0.24 - -0.07) the model significantly overestimated (observed/expected 0.85), the risk of infections or neoplasms in the external validation dataset. Also, the calibration slope was significantly lower than 1 (0.66, 95% CI 0.43-0.89), which means that the actual high risks are predicted to be even higher and the low risks are predicted to be even lower. Unlike the efficacy endpoints for which a regression was chosen as the final model, the effect of the treatment arm was not visible as clear separation in the predicted risk axis of the calibration plot. This was due to the facts that not only the coefficient of treatment in the final model for immunosuppressant safety was almost one tenth of that in the final models for relapse or new/enlarging lesions, but also, unlike the regression models for efficacy, there were predictors with greater coefficients than that of treatment in the final model for immunosuppressant safety.

The model predicting the composite outcome performed worse than in the development dataset (AUC at 24 months: 0.58, 95% CI 0.53-0.63) but the lower boundary of the 95% CI of its AUC(t) was above 0.50 for every monthly time point. Yet, the scaled Brier score at 24 months (0%) and its graph revealed lack of difference from a null model. The calibration-in-the-large (0.07, 95% CI -0.03 - 0.16) indicated that the model slightly but non-significantly underestimated (observed/expected 1.07) the risk of having the clinical efficacy or safety outcomes. A unit difference in the predicted risk corresponded to less than half unit difference in actual risk, as revealed by the calibration slope (0.46, 95% CI 0.18-0.74).

4.4.2 Decision and treatment response analyses

The decision curve analysis in the external validation dataset revealed that basing decisions on prognostic predictions from the relapse model would be informative between the risk thresholds of approximately 20 to 50% at 24 months (**Figure 12**). All patients willing to take the risk of an intervention, to avoid experiencing a relapse with 20% probability or less should be intervened with and no patients that would require at least 50% probability of relapse to take the risk of an intervention should be given one. Visual inspection of treatment effect modification by the treatment effect curves (**Figure 13**) that display predicted risk separately for the treatment arms, and the distribution of predicted treatment response revealed that daily treatment with fingolimod 0.5 mg was predicted to be superior to placebo across the board (**Figure 13**). All participants in the external validation dataset would be recommended fingolimod if the decision threshold was having any predicted benefit over placebo (Appendix B). This indicates that there was no qualitative heterogeneity in relapse risk in response to fingolimod, as expected by the lack of interaction terms in the final prediction model. The curves of risk conditional on treatment effect and their separate confidence intervals revealed very slight quantitative treatment heterogeneity, in which the risk of event with fingolimod 0.5 mg seem to increase less than with placebo as one covers the portions of the population with greater predicted treatment response. The empirical estimation of average benefit of fingolimod in those recommended was 0.22 (95% CI 0.15-0.30). The

predicted treatment response varies very narrowly (0.001) around the mean. The total gain from treatment response predictions over the marginal effect of fingolimod was the lowest for the relapse outcome (2.9%).

The 24-month risk threshold range for which the model predicting new or enlarging lesions in T2 MRI was useful was wide but high (approximately 40 to 90%). If the risk tolerance is outside this range, blanket strategies of intervening to all or none should suffice. Similar to the relapse outcome, the variability of predicted treatment response on new or enlarging T2 MRI lesions was very low (0.001) and all patients would be recommended treatment with fingolimod. So, there was lack of observable heterogeneity in response to fingolimod, as indicated by non-overlapping curves of new or enlarging lesion risk given treatment effect and low total gain from treatment response predictions over the marginal treatment effect (3.4%). The empirical estimation for the average benefit of fingolimod in those recommended was 0.26 (95% CI 0.19-0.33).

Although net reduction in interventions for the outcome of CDP at 24 months looked greater than that of the other efficacy outcomes, this was not due to a greater range of thresholds for which the prediction model was useful but rather due to the low incidence of the event, which made the strategy of no intervention a viable option. The model predictions for CDP were useful in a narrow risk range between approximately 25 to 35%. The mean predicted response to fingolimod was barely distinguishable from null and there was a qualitative heterogeneity of treatment effect where 52% (95% CI 48-56%) of the participants would be recommended daily fingolimod 0.5 mg as opposed to placebo according to the model predictions of CDP. There was some variability in predicted response to fingolimod (0.006). However, overlapping confidence intervals of the risk curves for CDP given treatment effect indicates that one cannot conclude a significant heterogeneity in treatment response. The total gain from model predictions compared to the average treatment effect was 6%.

As expected from its lack of discriminatory power, the model for predicting safety risk at 24 months did not have any benefit over the blanket strategies of intervention to all or none. The predicted response to fingolimod had some variability (0.009) around null and there was a non-significant qualitative heterogeneity of treatment effect where 49% (95% CI 45-53%) would be recommended treatment based on model predictions. The total gain from predicting the treatment response with the model was 8% but this descriptive statistic should not lead to the conclusion that the model is useful just because its predictions reveal heterogeneous treatment response. The absolute model predictions for the safety risk are expected to be incorrect, as evidenced by the lack of discriminative power.

The incidence of serious and non-serious infections and neoplasms was high in both the active and control treatment arms of the trials repurposed for this study. Hence, preventive measures would be recommended up to the threshold of approximately 75% event risk tolerance. For risk tolerance higher than 75%, the prediction model would be useful for decisions. The predicted treatment response distribution for the immunosuppressant safety outcome had some variability (0.005) and there was a non-significant qualitative heterogeneity where 49% (95% CI 45-53%) of the participants would be recommended fingolimod. The total gain from predicting the treatment response with the model as opposed to using average treatment effect was 6%.

Finally, the composite of safety and efficacy had a narrow range of threshold probabilities where the model was useful: 55-65%. The interpretation of the decisions of intervention and evaluating a threshold is complicated due to its composite nature. The predicted response to fingolimod varied asymmetrically around its mean (0.19, range 0.1-0.45) and its variance was 0.012. The outcome risk given treatment showed significant interaction with the predicted treatment effect. The total gain from predicting the treatment response was 10% (95% CI 3%-18%), the highest among all outcomes.

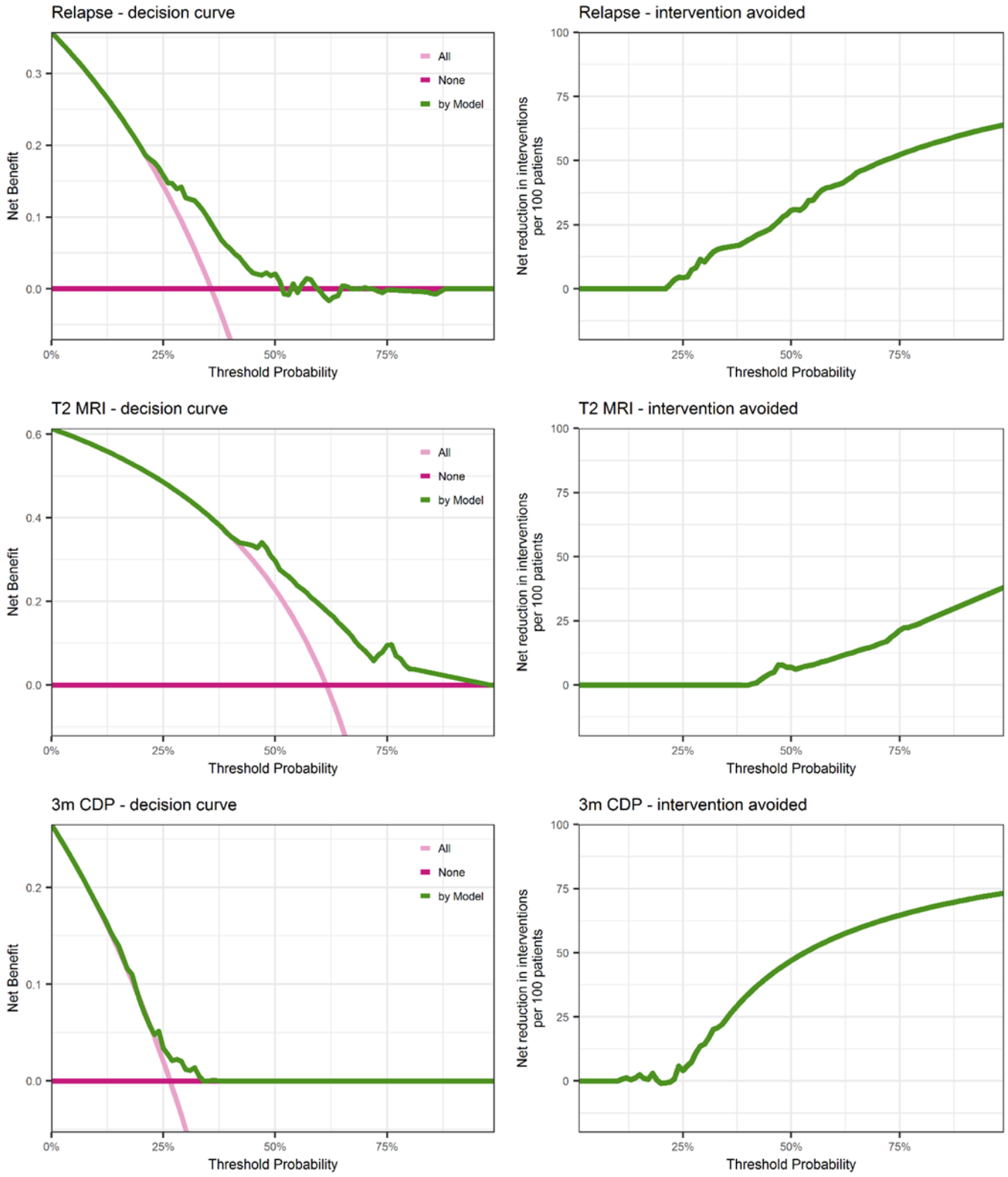
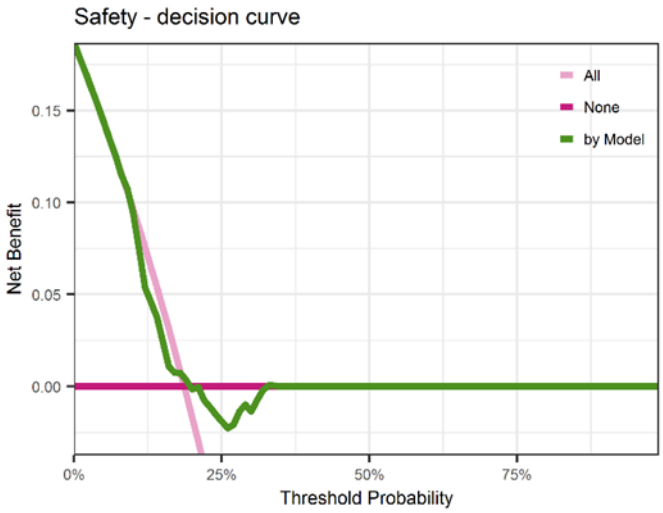
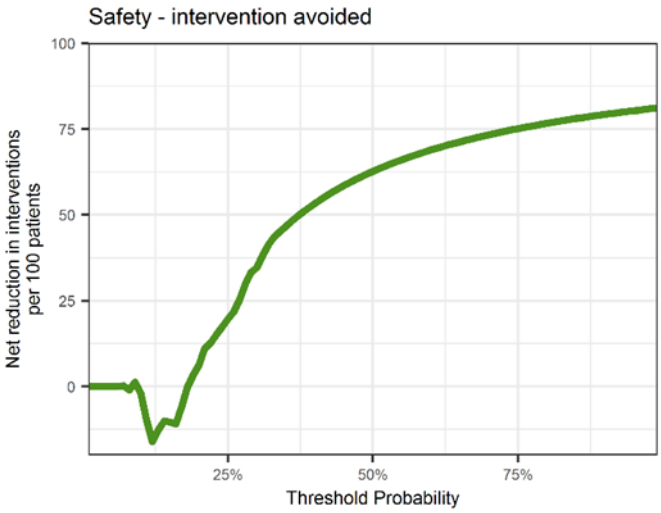
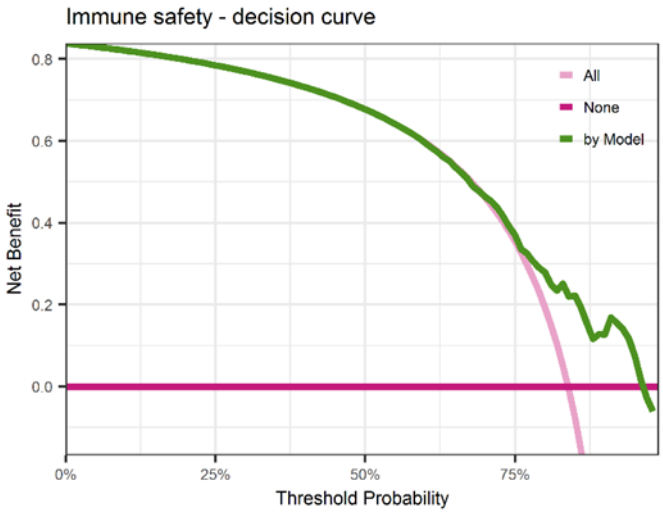


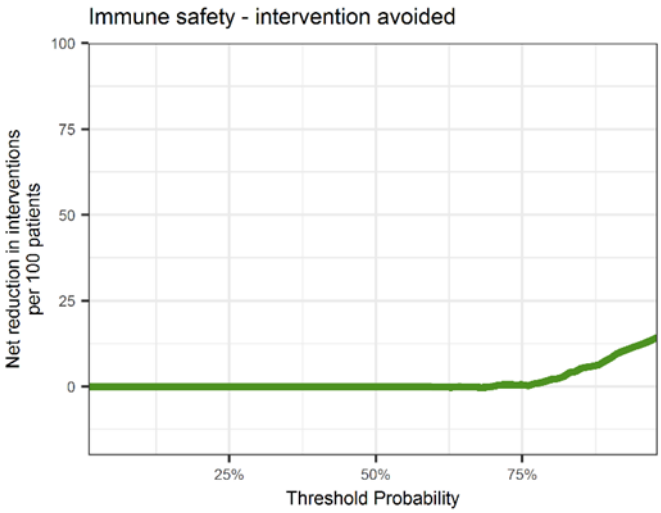
Figure 12 a/b *Decision curve analysis*

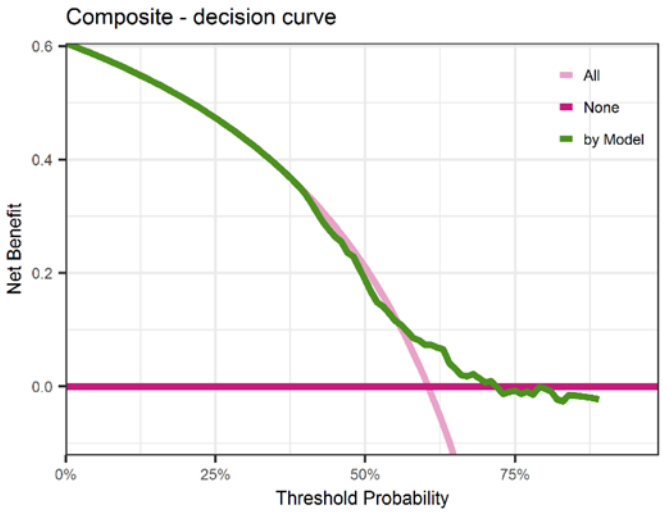
Expected net benefit under different strategies of intervention to all, none, or by model (left), and percent reduction in interventions using the model (right) for different risk thresholds **a)** *This page*: Relapse, New/enlarging lesions (T2 MRI), and and Confirmed disability progression (3m CDP) **b)** *Next page*: Safety, Immunosuppressant safety (Immune safety), and Safety and efficacy (Composite).

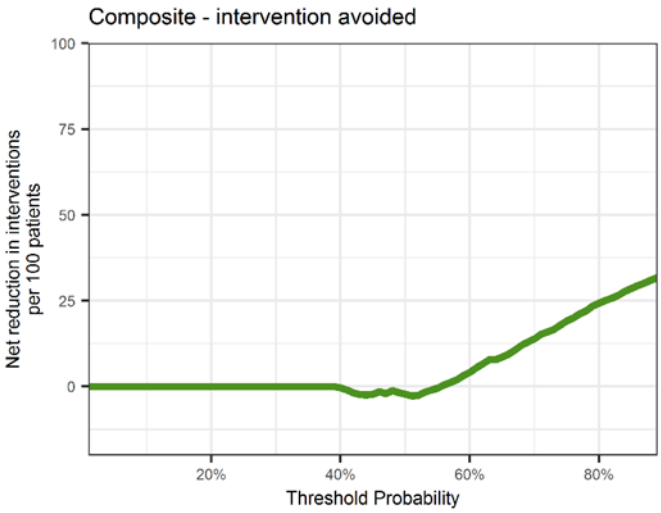

Safety - decision curve
All
None
by Model
Net Benefit
0.15
0.10
0.05
0.00
0%
25%
50%
75%
Threshold Probability


Safety - intervention avoided
Net reduction in interventions per 100 patients
100
75
50
25
0
25%
50%
75%
Threshold Probability


Immune safety - decision curve
All
None
by Model
Net Benefit
0.8
0.6
0.4
0.2
0.0
0%
25%
50%
75%
Threshold Probability


Immune safety - intervention avoided
Net reduction in interventions per 100 patients
100
75
50
25
0
25%
50%
75%
Threshold Probability


Composite - decision curve
All
None
by Model
Net Benefit
0.6
0.4
0.2
0.0
0%
25%
50%
75%
Threshold Probability


Composite - intervention avoided
Net reduction in interventions per 100 patients
100
75
50
25
0
20%
40%
60%
80%
Threshold Probability

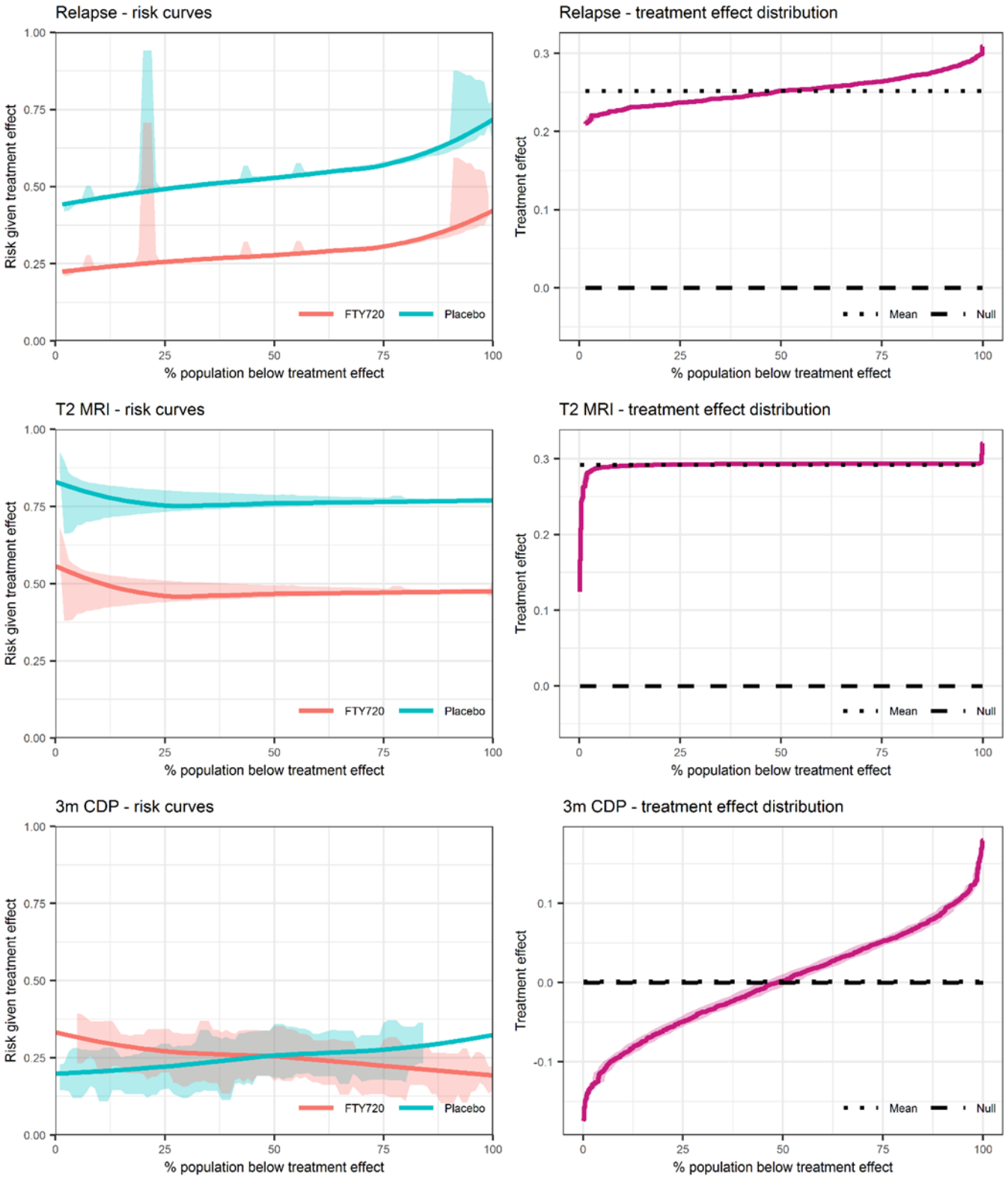
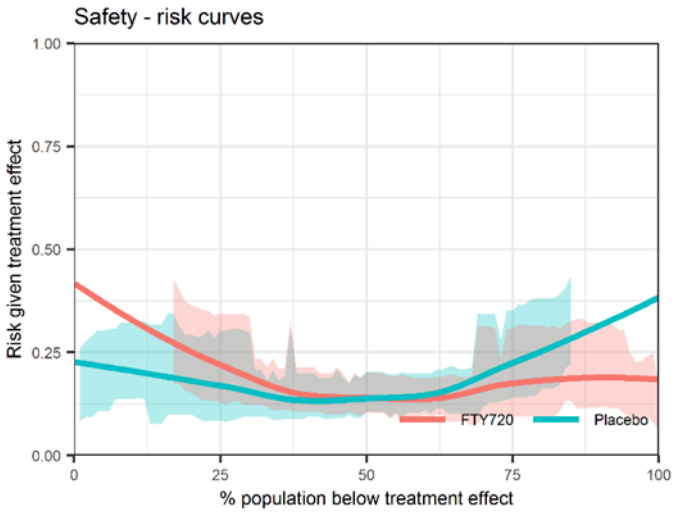
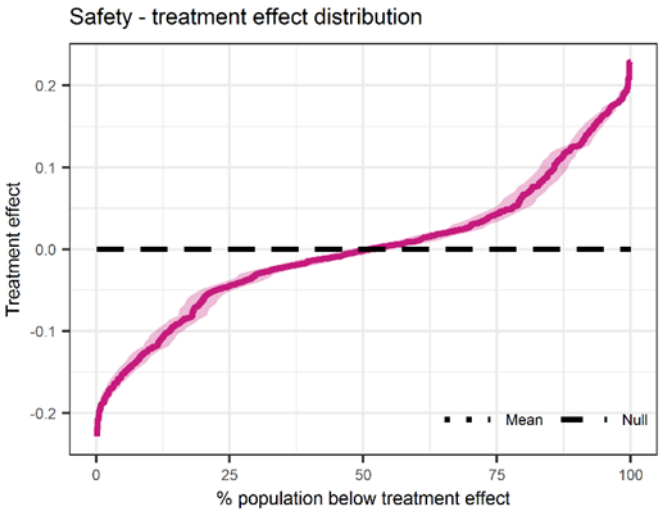
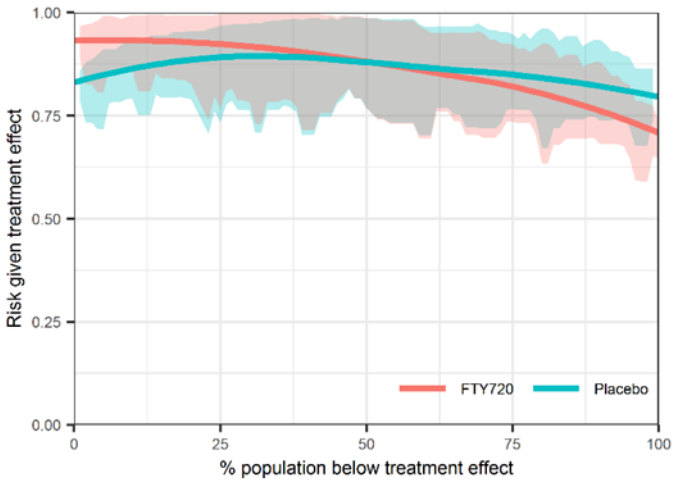


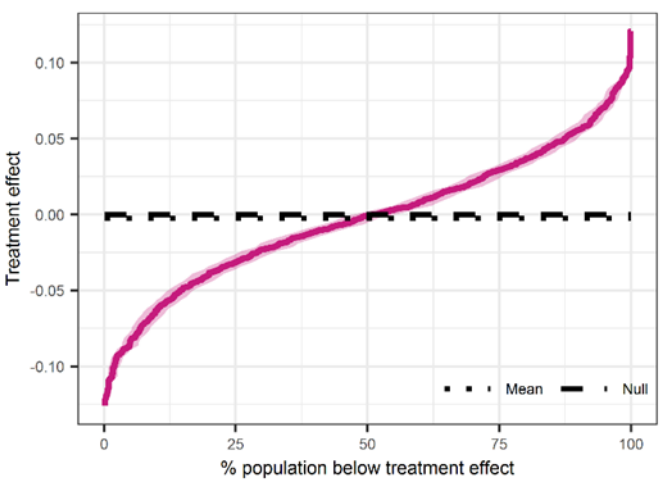
Figure 13 a/b *Predicted treatment response*

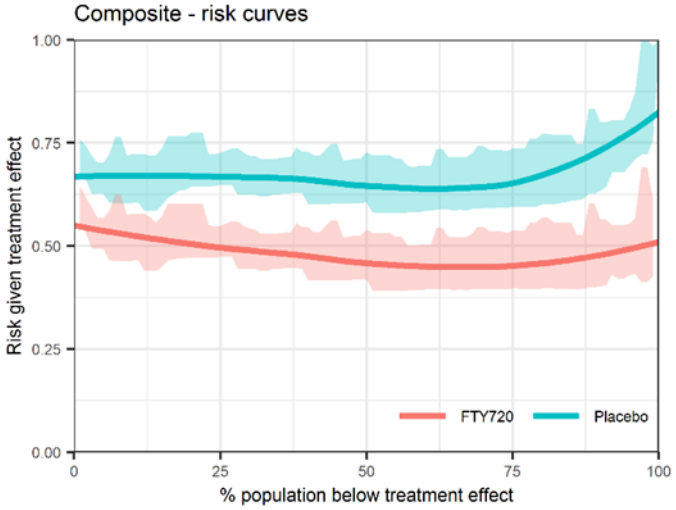
Treatment effect prediction in the external validation dataset. Predicted risk under each treatment to investigate effect modification on the left, and distribution of predicted treatment effect on the right, **a)** *This page*: Relapse, New/enlarging lesions (T2 MRI), and Confirmed disability progression (3m CDP) **b)** *Next page*: Safety, Immunosuppressant safety (Immune safety), and Safety and efficacy (Composite).

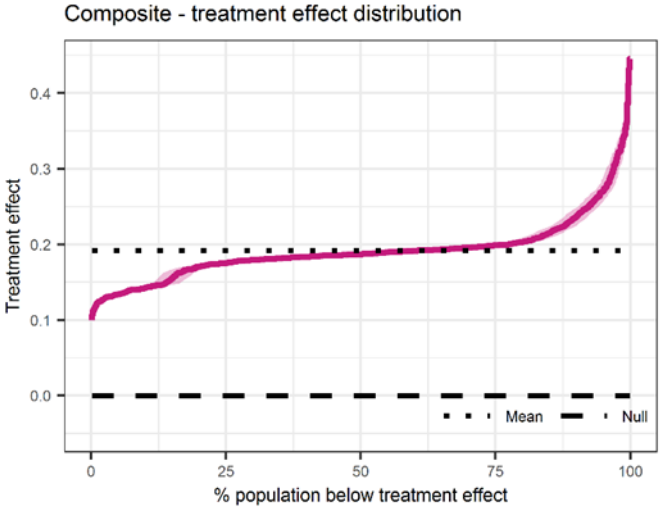

Safety - risk curves
Risk given treatment effect
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
% population below treatment effect
FTY720
Placebo


Safety - treatment effect distribution
Treatment effect
0.2
0.1
0.0
-0.1
-0.2
0
25
50
75
100
% population below treatment effect
Mean
Null


Immune safety - risk curves
Risk given treatment effect
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
% population below treatment effect
FTY720
Placebo


Immune safety - treatment effect distribution
Treatment effect
0.10
0.05
0.00
-0.05
-0.10
0
25
50
75
100
% population below treatment effect
Mean
Null


Composite - risk curves
Risk given treatment effect
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
% population below treatment effect
FTY720
Placebo


Composite - treatment effect distribution
Treatment effect
0.4
0.3
0.2
0.1
0.0
0
25
50
75
100
% population below treatment effect
Mean
Null

5. Discussion

This Chapter starts with reviewing the final model, its performance, and its comparison to other prediction models in the literature, initially for the primary outcome of time-to-relapse (Section 5.1), then for the rest of the outcomes (Section 5.2). Influential predictors per outcome are examined and put in the context of the current knowledge in Section 5.3. The strengths and limitations of this work, alongside their reasons and possible effects, are evaluated in Section 5.4. How the results from this thesis can have an impact on future scientific inquires or clinical decisions and which inspirational goals remain unaccomplished are discussed in Section 5.5.

5.1 Predicting relapse

The final model for predicting the relapse risk was parsimonious with only main terms selected by a lasso penalty. Discrimination performance in the external validation dataset measured by AUC at 24 months was 0.68 (95% CI: 0.63-0.72), which was very close to the AUC estimated via cross-validation in the model development dataset. This performance is very similar to the internally validated *c*-statistics of 0.65[136], of 0.62[122], and of 0.65[140] for models predicting relapse in similar studies. Although the discriminative performance of the final model in this thesis was moderate and significantly higher than 0.5, the Brier score in the external validation dataset was not much better than that of the null model and the calibration assessment revealed overestimation of risk (observed/expected 0.84). Before application to new patients from other populations, the model may need recalibration. Basing interventional decisions on the relapse model seemed to be useful when the threshold relevant for decision making was between 20 and 50% event risk at 24 months, covering almost three-fourths of the predicted event probabilities.

Luckily, for the primary outcome of relapse, a simple, robust, and easy to report model was selected via cross-validation. Although they may be more flexible, complex black-box algorithms like the random forest are more difficult to interpret and report.[188] The best model predicting relapse was a regression rather than a tree-based method and no predictor by treatment interactions were selected in the final model. These indicate a lack of detectable heterogeneity in treatment response, at least as predictable by the commonly measured variables used in this study. Also, according to the counterfactual predictions and the assumption of no cost or harm, all participants in the external validation dataset would benefit from receiving daily fingolimod 0.5 mg rather than placebo. There was no qualitative treatment effect heterogeneity and only a very low variability of 0.001 in predicted treatment response. Hence, the individualized treatment response predictions by the model did not result in a relevant gain (2.9%) over the marginal treatment effect. The individual risk of relapse at 24 months was predicted to be lower with fingolimod for all, with predicted absolute risk reduction of 21 to 31% with fingolimod compared to placebo.

The findings from other comparable studies predicting heterogeneity in treatment response for the relapse risk similarly point to possible lack of detectable or predictable qualitative heterogeneity, not only for fingolimod but for other marketed DMTs. Chalkou and colleagues[136] reported that although the main terms for the DMTs they investigated (natalizumab, glatiramer acetate, dimethyl fumarate vs. placebo) and the prognostic risk score they developed were statistically significant in the meta-regression, the DMT by risk score interaction was not. Also, the study by Stühler and colleagues[140], which included fingolimod alongside other five DMTs, revealed less than 0.01 difference in cross-validated *c*-statistic

from the prognostic model containing only main terms and the predictive model containing both main and treatment interaction terms. In contrast, Pellegrini and colleagues[139] reported qualitative heterogeneity of response to dimethyl fumarate compared to placebo based on statistically significant interaction of treatment with the treatment response score they developed. This may be due to the fact that Pellegrini and colleagues[139] chose the model that maximized the treatment effect heterogeneity as opposed to models by expert-view or by maximizing prognostic model fit. It remains unclear whether their treatment response model can be converted to a well-calibrated absolute risk that may be useful for individual decision-making.

5.2 Predicting other outcomes

The final model predicting new or enlarging T2 MRI lesions was developed with an elastic net penalty and only had main term effects. The 24 month AUC was 0.74 (95% CI: 0.70-0.78) at external validation and it was the highest for this outcome among others investigated in this study. This discrimination is much better than the cross-validated AUC of 0.62 from the model with a similar outcome.[151] This discrepancy can be attributed to the difference in data source (RCT in this study vs. routine care by Hapfelmeier), the difference in the treatment included (fingolimod vs. first-line), or the difference in the outcome conceptualization (right- vs. interval-censored). The good fit of the final model in this study was also demonstrated by the highest scaled Brier score (12%) at 24 months. The calibration performance of the final model predicting new or enlarging T2 MRI lesions was moderate because the calibration plot was scattered around the diagonal when stratified by drug, precluding strong calibration.[189] The model would be useful in decision making when the risk threshold is between 40 and 90%. Results from the treatment effect analysis in the external validation dataset for new or enlarging T2 MRI lesions was similar to that for relapse. In short, daily fingolimod 0.5 mg would be recommended to all participants, and there was very low variability in predicted treatment response with very low gain from model predictions. Although this non-clinical efficacy outcome was easier to predict with the existing predictors, there was no evidence of treatment effect heterogeneity. Also, it is unclear how this predictability can be translated into the clinic for decision making during which clinical outcomes are of primary interest.

With a cross-validated $AUC_{avg}(t)$ of 0.67 for the 3-month CDP outcome, the transformation forest outperformed other methods by over 0.1 difference, which may be indicative of the need to include non-linear or higher-order interactions to represent this disability endpoint. This discriminatory performance was comparable to the bootstrapped *c*-statistic of less than 0.65 for the prognostic model predicting disability progression[133], cross-validated AUC of less than 0.69 for the prognostic model predicting CDP[134,135], but better than the cross-validated *c*-statistic of 0.56 for the predictive CDP model[140] in similar studies. The external validation revealed worse discriminatory performance for the CDP model developed in this study, especially as the prediction timeframe got longer, with an AUC of 0.59 (95% CI 0.54-0.64) at 24-months. Also, the Brier score of the final model was very close to that of the null, making the quality of predictions from the model questionable. Although there seemed to be qualitative heterogeneity of treatment effect, as evidenced by 0.006 variance of predicted treatment response around null, the difference in risk given treatment was never significant between treatment arms. The null median predicted treatment response was unsurprising in the light of the reported results from the source trial, FREEDOMS II, in which fingolimod 0.5 mg failed to have a significant effect on the risk of 3-month CDP.[55]

In this study, a novelty was the aim to develop prognostic models to predict safety-related outcomes. The safety outcome, composed of any SAE or discontinuation of the trial due to an adverse event, could not be modeled with sufficient discriminatory power. The best performing method was transformation

forest and had a cross-validated AUC of 0.54. In the external validation dataset, the final model had an AUC of 0.50 at 24 months. This result points to the fundamental difficulties in modeling safety outcomes. If adverse events from multiple SOCs are pooled, the heterogeneity makes it harder to capture their different underlying mechanisms in a model. Yet, adverse events, especially serious ones, are expected to be not very common with marketed drugs. As one increases the granularity of grouping adverse events to better predict them, there are fewer events in each group and power decreases. Combining results from multiple safety-related models is an unaddressed methodological challenge.

The more refined outcome of immunosuppressant safety, which is related to the main mechanism of action of all DMTs, including fingolimod, was more predictable than the overall safety outcome. The best performing method to predict an adverse event from the SOCs of infections and infestations or neoplasms was elastic net. It had moderate cross-validated performance with an $AUC_{avg}(t)$ of 0.60 and external validation performance with an AUC of 0.69 (95% CI 0.63-0.74) at 24 months. However, the overall fit and calibration performance in the external validation dataset were not satisfactory. There was some non-significant heterogeneity in the predicted treatment effect on immunosuppressant safety demonstrated by the variability of 0.005 and 6% gain from the predictions over the marginal treatment effect.

Even though the eventual goal of personalized medicine is estimation of the net benefit that an individual expects, combining the results from separate models that predict efficacy and safety outcomes is a challenge.[94] Forming a composite outcome as was done in this study is far from the ideal solution. It assumes that all events contributing to the composite outcome have the same weight or value and that a single model is feasible despite potentially different biological mechanisms behind them. Grouped lasso was the selected method to predict the outcome formed as a composite from relapse, 3-month CDP, and safety. The model had low discriminatory performance of AUC 0.58 (95% CI 0.53-0.63) at 24 months in the external validation dataset. The overall fit was not different from a null model, according to the Brier score improvement of 0%. The predicted treatment response for the composite outcome was the most heterogeneous, as indicated by its variance of 0.012 and high total gain of 10% from predicting the response to fingolimod compared to using its marginal effect. The interpretation of the results and model usefulness is challenging for composite outcomes.[33]

5.3 Important predictors

The variable that had the greatest influence in the final model predicting time-to relapse was the drug. Being assigned to the active treatment arm of daily fingolimod 0.5 mg, as opposed to placebo, increased time-to relapse. In addition to the drug, only four predictors were selected by the lasso penalty in the final model and all were main terms. These had the same direction of effect and were also selected by the logistic regression with lasso penalty in the risk score developed by Chalkou and colleagues.[122] Included in the list were established predictors indicative of disability, as total EDSS score, and symptoms, as number of relapses in the last 2 years, which were used as adjustment factors in the primary analysis of ARR in the source trials of FREEDOMS and FREEDOMS II. Higher disability or disease activity, as demonstrated by baseline EDSS or pre-baseline number of relapses, increased the risk of and decreased the time-to first relapse, confirming the results from the trials' conventional subgroup analysis.[65] Another predictor selected in the final model and deemed important by all modeling methods was total volume of Gd-enhanced T1 lesions, an MRI-related marker considered to detect new inflammatory activity[19] and surrogate of ARR for many DMTs.[36] Lasso penalty is expected to select only a single one of correlated variables. Hence, the selection of total volume of Gd-enhanced T1 lesions may hint that it may be capturing something beyond the pre-baseline number of relapses. As expected

from the results of the trials' conventional subgroup analyses, higher volume of Gd-enhanced T1 lesions increased the risk of relapse.[65] Also confirming the results from the subgroup analyses, the last variable that predicted time-to relapse in the final model was the number of prior MS treatments. Exposure to higher number of DMTs before the trial meant higher risk of experiencing a relapse during the trial. The mechanism by which these two are related is not easy to disentangle. Maybe the number of prior MS treatments reflects highly active disease beyond the number of relapses, EDSS, and MRI measures. Or, maybe, the residual rebound effect from previous treatments lasts longer than the wash-out period of 3 to 6 months as foreseen by the eligibility criteria of the source trials. Unsurprisingly, another surrogate MRI marker of ARR, total volume of T2 lesions[36] was important in predicting time-to relapse across methods during cross-validation, although it was not selected in the final model. Interestingly, a predictor deemed important for relapse risk across methods was the concomitant diseases from the SOC of metabolism and nutrition disorders. One possible explanation is that the bioavailability of fingolimod (and hence its comparator placebo) may be different in those with metabolic disorders.[190] Another explanation may be an unobserved underlying mechanism that effect both experiencing metabolic or nutritional disorders and future relapse risk. Age or sex, which were significant in subgroup analyses of fingolimod trials[64,65], were not found to be influential independent predictors by the methods employed in this study.

In line with the results from the source trials[54,55], daily fingolimod 0.5 mg decreased the risk of experiencing new or enlarged T2 MRI lesions compared to placebo in the final prediction model. The drug was the most influential variable followed by, unsurprisingly, other MRI markers: total volume of Gd-enhanced T1 lesions, number of Gd-enhanced T1 lesions, and total volume of T2 lesions. Like a self-fulfilling prophecy, higher volume or number of MRI lesions at baseline predicted shorter time-to new or enlarged T2 lesions. Age categories and disease duration were also in the final model. Other important but unexpected variables that were selected by the elastic net penalty in the final model were visual analog scale (QoL), and bilirubin, increase of which increased the risk of new or enlarging T2 lesions.

For predicting 3-month CDP, the best method was a transformation forest indicating that probably some higher order interactions and drug by predictor interactions were relevant. Mean of 9HPT from the MSFC panel, and concomitant diseases from the SOC of musculoskeletal and connective tissue disorders were deemed important by three modeling methods during cross-validation. Because CDP is a disability related outcome measured by EDSS, it is unsurprising that according to the finally selected method of transformation forest, total EDSS score and EDSS system score of the cerebral (or mental) functions were important predictors in at least two of the cross-validation folds. The same argument can be made for the 9HPT measuring hand dexterity.[191] When the baseline EDSS score is greater than 4.0, the main determinant of its change is ambulatory functions[33], which may be a reason for concomitant musculoskeletal and connective tissue disorders to predict CDP. These four predictors found important in this study do not intersect with those reported by other comparable prediction studies with disability outcomes.[133,138] There may be many reasons for such a discrepancy. Primarily, interaction with fingolimod treatment in a global outcome prediction framework via a model-based tree is unique to this study, whereas Bovis and colleagues[138] specifically modeled CDP as a response to another DMT, and Pellegrini and colleagues[133] modeled prognosis of a composite measure of disability in only placebo arms by multiple methods, including support vector machines. Either these methodological differences caused non-overlapping lists of important predictors, or there are many but weak (or differently measured) predictors of (differently measured) disability increasing the randomness of which ones are found important in different studies. Similar to this study, the pre-print by De Brouwer and colleagues[142] also reports EDSS and functional system scores as high ranking in importance for predicting CDP. Especially worth noting in this study is the insignificance of disease activity indicators like MRI markers

for predicting 3-month CDP, compared to the high influence of MRI lesions on relapse and new or enlarging T2 MRI lesions. This observation is contradictory to the conventional subgroup analyses of the FREEDOMS trial[64], in which number of relapses and T2 lesion volume was significantly interacting with treatment. Yet, results from this study is in line with the so called “clinico-radiological paradox” in MS[34] and raises the question whether disease activity and disability progression may have different mechanisms. Other analyses based on trials of the same drug had similarly conflicting results.[192]

The single predictor deemed important by all methods for predicting the safety outcome was the concomitant diseases from the SOC of gastrointestinal disorders. Hypothetically, the oral route of fingolimod’s administration may be the underlying mechanism by which patients with concomitant gastrointestinal disorders have different bioavailability of the drug and hence the risk of experiencing SAEs or discontinuing the study due to an adverse event. In predicting immunosuppressant safety, the predictor found important by three modeling methods was exposure to comedications of genito urinary system and sex hormones as important to predict risk of infections or neoplasms. Unless there is a drug interaction between fingolimod and this class of comedications, it is unclear how this predictor may be prognostic.

5.4 Strengths and limitations

Conducting this study was only possible because the sponsor of these trials, Novartis, put in place mechanisms and accepted our research proposal to share their data. Novartis has established further initiatives for advanced data sharing in the disease area of MS.[193] Experts in the field of treatment effect prediction have been calling for data pooling opportunities[94] because identification and prediction of treatment effect modification via multivariable modelling techniques require high sample sizes for sufficient power. In this study, data from the trials of a single drug were used not only to decrease the complexity of the prediction task but also to be able to conveniently investigate candidate predictors, like QoL, measured consistently with the same tool by a single sponsor. This study evaluated a single treatment option for MS patients although there are many. Forming a large randomized dataset compatible for prediction modeling is a challenge because the baseline characteristics collected and how they are recoded vary greatly between trials.

The data source is both a limitation and a strength of this study. Compared to the patients that are encountered in routine clinical practice, RCT participants tend to be more homogenous due to the eligibility criteria.[68] In the context of MS, RCT participants tend to have higher disease activity and less comorbidities.[123] Even though patient selection into the trials may have hindered identification of some variables as predictors or forming a prediction model closer to the real world, randomization provides an unbiased way to predict heterogeneity in treatment response. In contrast, non-randomized settings would entail confounding by indication.[67,145] Also, data quality and completeness is expected to be superior to any observational data that was collected for purposes other than prognostic prediction. Such secondary use of data is relatively common in MS prognostic prediction literature.[129]

Regardless of being experimental or observational, secondary use of data brings challenges to prognostic modelling. Not all candidate prognostic factors are necessarily measured when the primary aim of data collection is different. In the source trials of FREEDOMS and FREEDOMS II, no CSF measurements were available. The range of measured MRI-related variables were readily available in routine practice, yet limited. For instance, measures on brain atrophy, which are thought to be related to disability progression[123], could not be included as candidate predictors because at baseline there was only absolute brain volume but no % change over a period of time. Other MRI measures available at specialty centers or results from –omics analysis were not available either, although one can argue

against their inclusion anyways because these would make a prognostic prediction model difficult to use or implement in routine care. Although they are prognosis-wise promising, CSF or genetic markers have unfortunately not been investigated in well-conducted and sufficiently powered prediction modeling studies in MS.

The limitations of the source trials are likely to inflict this study, too. According to a systematic review, the different reasons for drop-out in the arms of the source trials FREEDOMS and FREEDOMS II increased their risk of bias.[58] This may have caused misspecification of the model. Also, if the discontinuations were related to the outcome, censoring may have been informative. Two years is a common timeframe for pivotal trials in MS. Yet, due to the chronicity of this lifelong disease and longer time required to observe disability-related outcomes, it may be considered short for prognostic purposes.[123] What would happen after 2 years and whether the developed models would have predictive power beyond this timeframe could not be answered in this study.[67]

Another strength of this study due to its data source is consideration of predictors that were not previously investigated. Because comedications and concomitant diseases are systematically recorded in RCTs, including the source trials of this study, groups of these could be investigated as candidate predictors of efficacy and safety outcomes. The important groups can further be investigated in future studies to identify which diseases or comedications, among many, can be confirmed to have predictive power.

The participants randomized to the high dose of daily fingolimod 1.25 mg in the source trials were excluded from this study. Due to the randomized nature of the treatment assignment, it is unlikely that this decision caused any selection bias. On the contrary, the models developed are more fit to be further evaluated in real-world datasets because only the dosage available on the market was used.

EPV in the model development dataset (1.9 for the primary outcome of relapse) was lower than 10, which is the minimum recommended value in the prognostic modeling literature.[194] However, three of the four competing modeling methods had shrinkage either due to a penalty factor, in the case of regularized regressions, or due to an averaging over an ensemble, in the case of random forest. Shrinkage is expected to minimize overfitting to the training dataset, a problem exacerbated by low EPV. Eventually, models from these methods were chosen to have the best performance. The satisfactory performance of the model predicting relapse in the external validation dataset also makes the problem of low EPV in the development dataset less critical. The effective sample size (number of events) in the external validation dataset was satisfactory for relapse and other secondary endpoints.

The outcomes of CDP and new or enlarging T2 MRI lesions, which are only observable during regular visits, could have been more correctly conceptualized as interval-censored rather than right-censored. However, modeling and evaluation methods for this type of outcome are either limited or non-existent. Because the optimization and evaluation process would become very different and these two outcomes were considered secondary, right-censored analysis was sought for all investigated outcomes.

The two unverified assumptions for all the modeling methods were missing at random, based on which missing was imputed, and non-informative censoring, based on which time-to-event methods were used. If violated, the assumption about missing data is not expected to cause a major change in results because of the very low frequency of missing per predictor with median 0% and maximum 6.5% in model development dataset, and median 0% and maximum 1.8% in the external validation dataset. However, informative censoring due to missing outcome measurement could potentially distort the developed models or performance evaluations because, respectively, 13% and 21% of participants in the model development and external validation datasets were missing any 24-month visit.

Not investigating the assumptions of the modeling methods is not necessarily a limitation. When the objective is prediction rather than inference, the modeling assumptions become less relevant as long as the prediction performance is satisfactory.[84] The investigation of and correction for any violated modeling assumptions are valuable only if predictions become more accurate. The assumption of proportional hazards in the penalized regression models was indirectly challenged by including an alternative competing method, the transformation forest, with the ability to detect non-proportional hazards.[162]

The choice of competing modeling methods determines the possible functional forms that can be represented. In this study only generalized linear regressions and tree structures were investigated. This does not preclude other possible functional forms that may be better fitting to the data, e.g. support vector machines. Higher order terms or interactions between predictors, except those with treatment, were not included in the penalized regressions, assuming an additive, log-linear relationship between the predictors and the outcome.[84] Yet, if this assumption was strongly violated, the random forest – being a method that is good at capturing higher-order relationships - would be expected to outperform the penalized regressions. It was the case for the CDP and safety outcomes but not others. Also, treatment interaction effects are expected to be weaker than prognostic ones and could be handled differently in the penalized regression.[45,187] In this study, if the weaker treatment interaction effects were relevant to the prognostic performance, the grouped lasso or recursive partitioning methods would be expected to outperform elastic net because the discriminative power of prediction models is expected to decrease when true interactions are omitted.[78]

The cross-validated discrimination performance of the transformation tree varied very close to 0.50 for all investigated outcomes in this study, indicating that none of the outcomes could be predicted well by the structure of a single decision tree. Similarly, a systematic review of 71 prediction modeling studies with competing methods found that when included as an alternative, trees always had worse performance than other machine learning methods. Authors of that systematic review concluded that in comparative studies with low risk of bias, machine learning methods on average did not outperform linear regression based methods in terms of the validated AUC.[85] Their conclusion is in line with the results from this study.

There are many ways of modeling differential treatment effects, statistical methods of which have been developed recently.[45] These could well be applied to answer the questions posed in this study, too. However, head-to-head comparison of the performance of all these methods is missing. A study compared five of them in terms of their performance in recovering the underlying structure of simulated data and found that model-based recursive partitioning, the method on which transformation forests included in this study are based, was the best at detecting treatment predictor interactions under varying conditions compared to the others.[195]

The well-known barriers to predicting individualized treatment response are also the limitations of this study. Evaluation of the predicted treatment response is constrained by the fact that it cannot be observed. The lack of known strong effect modifiers precludes selection of handful of predictors to include in the model based on medical expertise. This is compounded by the fact that RCTs are expected to be underpowered to detect multiple weak interactions in a modeling framework.[103] Based on the results for the primary outcome of relapse, it is questionable whether the response to fingolimod is heterogeneous enough to require individualized predictions and, if it is, whether this can be predicted based on the collected variables.[196] The deterministic assumption that the variability in the outcomes observed in the active arm in an RCT is necessarily a quality of the patient and cannot vary within the patient in a random manner is a strong one. In terms of prediction, a correct model generally does not exist and there may be subtle and higher term interactions, or unobservable or yet to be discovered predictors.[84]

In this study, the probabilistic predictions from the prognostic models were intentionally not categorized to create a decision rule or to define risk groups. This action should be medically justified and depend on weighing of benefits and harms of the outcomes, either by healthcare professionals or by patient representatives.[84]

5.5 Implications

The translation of a prediction model into clinical practice requires many stages, the initial of which was addressed in this study (**Table 15**). The primary aim was to develop a model predicting risk of relapse and to assess its reproducibility in an independent but similar group of patients. The result was a simple prognostic model with a moderate AUC of 0.68 in new patients at 24 months. The model can predict relapse risk in RRMS patients receiving no treatment (placebo) or fingolimod at time points up to 2 years. If the decision thresholds at which this model has high net benefit are considered clinically relevant by medical experts, it can be further externally validated, preferably in observational datasets that are more representative of patients encountered in routine care. A well-performing model can be useless in the clinic, whereas a poor-performing model very useful[117], if predictions can help optimize the treatment procedures. Updating the model was not sought in this study. But, as indicated by the overestimation in the external validation dataset, recalibration may be necessary in a different clinical setting that the model is planned to be used.[84] After demonstration of transportability by additional satisfactory external validations[110,197], the model can be turned into a tool for impact analysis[198] and implemented in healthcare settings.[146,199]

Unless the harms or costs of treatment are taken into account, the predictions from the relapse model suggested treating all RRMS patients in the external validation set with daily fingolimod 0.5 mg to maximize the treatment effect both at the individual and the population levels. This may change when burdens are reflected in the threshold for accepting a treatment. The decision threshold is a medical question rather than a statistical one and should be addressed taking into account the purpose and context of using the treatment effect model.[75] Treatment response prediction models differentiating patients that would benefit from a treatment as opposed to a comparator are only expected to be useful when the clinically relevant decision threshold range is close to the expected outcome from the treatment, in which case the model would separate those that benefit from those that are harmed. In the case of the relapse risk model, the range of the predicted individual response to fingolimod was between 0.21 and 0.31 decrease in relapse risk at month 24. Whether this treatment effect range would be relevant in the clinic should be decided on a case-by-case basis. The goal of this study was not to find an algorithm for decision-making or replacing clinical judgement. The aim was to combine multiple prognostic factors and produce an estimated prediction of risk conditional on baseline factors, which may then be used to support treatment decisions.[198,200,201]

Lipkovich[45] suggests that qualitative interactions in which the experimental treatment is better than the standard treatment for a subgroup but worse for the rest is relatively rare. The lack of treatment interaction terms or interaction structure in the best performing model for predicting relapse risk at 24 months supports the idea that there may not be a heterogeneity in response to fingolimod beyond the heterogeneity in baseline prognostic risk in untreated RRMS patients. Even the heterogeneity in baseline risk introduced little variability to the predicted treatment effect. These observations are also valid for the outcome of new or enlarging T2 MRI lesions, the final model of which had properties and performance very close to that of relapse, highlighting the close relationship between clinical and imaging-based disease activity in RRMS.

Context	Output	Implication
Clinical, Research	Relapse model as a prognostic prediction tool	Implementation in clinics for use as a decision support tool during care of RRMS patients, only after: 1) decision thresholds of model benefit are found relevant by clinicians 2) further external validation in datasets from various sources 3) recalibration, if needed in a new setting 4) impact analysis
Clinical	Prediction of response to fingolimod versus placebo from the relapse model	• if ~21% reduction in two-year relapse risk in response to fingolimod outweighs its potential harms, then treat all RRMS patients* with fingolimod • if ~31% reduction in two-year relapse risk in response to fingolimod does not outweigh its potential harms, then do not treat similar RRMS patients* with fingolimod • if the decision threshold to treat with fingolimod lies between ~21 and 31% reduction in two-year relapse risk, calculate the predicted decrease in relapse risk with the relapse model
Research	Predictability and variability of response to fingolimod	Future studies in RRMS patients and using different data sources and candidate predictors: • to confirm or rebut lack of variability of clinical and imaging-based disease activity in response to fingolimod • to investigate the predictability of disability and safety, and variability of the change in their risk in response to fingolimod
Research	Important predictors	Future studies in RRMS patients and using different data sources to confirm or rebut the importance of: • concomitant metabolism and nutrition disorders for predicting relapse • concomitant musculoskeletal and connective tissue disorders for predicting CDP • bilirubin for predicting new or enlarging lesions in T2 MRI • concomitant gastrointestinal disorders for predicting overall safety • comedications of genito urinary system and sex hormones for predicting infections or neoplasms
Research	Further horizons for treatment response prediction in RRMS patients	• modeling the predicted treatment response as an outcome • modeling dynamic treatment regimens • combining treatment benefits and harms as a function and optimizing it after weighing of the outcomes

Table 15 *Implications*

Clinical and research implications from the results of this study. *RRMS patients similar to those included in the FREEDOMS and FREEDOMS II trials. RRMS: Relapsing-remitting multiple sclerosis, CDP: Confirmed disability progression, MRI: Magnetic resonance imaging.

The PATH statement suggests careful interpretation of analysis of treatment effect heterogeneity when there is no established overall treatment effect on the particular outcome to begin with because identification methods are more likely to lead to over-optimistic results.[78] This might be true for the 3-month CDP in this study, for which there may or may not be an overall treatment effect based on conflicting results from the source trials. There is no clear overall effect of treatment also on the safety-related outcomes in this study. Hence, the results on observed treatment effect heterogeneity for these outcomes should be considered with caution.

In light of the moderate prognostic performance of the final relapse prediction model in this and other well-conducted similar studies in RRMS patients[122,136,140], medium term relapse risk may not be highly predictable. The results for predicting disability are even more discouraging.[133,134,140] These results may indicate inherent non-predictability of these outcomes due to lack of true early predictors. A more optimistic interpretation might be that such true early predictors are yet to be identified. Future methodologically sound multivariable prognostic modeling studies in different data sources are likely to further shed light on this.

Via its secondary objective, this study addressed the gap in evaluating laboratory measurements, concomitant diseases, and comedications as predictors of outcomes relevant for RRMS patients. Apart from expected markers of disease activity and severity, concomitant metabolism and nutrition disorders, and concomitant musculoskeletal and connective tissue disorders were found to be important in predicting the efficacy outcomes of relapse and CDP, respectively. Also, the lab measurement bilirubin was influential in predicting new or enlarging T2 MRI lesions. Concomitant gastrointestinal disorders were important for predicting the overall safety outcome and comedications of genito urinary system and sex hormones for predicting infections or neoplasms. These results can only be considered exploratory and further hypothesis-driven research is warranted for the confirmation of their importance.[78,202]

There are various ways in which future work can add to what is put forward in this study. The predicted treatment response can become an outcome of its own and be modelled with different techniques[79], even though this approach would bring another layer of modeling and increase the chance of error. As an alternative to no drug use (or placebo use), only fingolimod treatment at a single time point was considered in this study. During the clinical decision-making process for RRMS patients, there are many more alternative treatments and time points for decision. These can only be addressed in large heterogeneous datasets, which are likely to be observational and to bring potential biases. Still, future work on dynamic treatment regimens for a chronic condition like RRMS would be valuable.[45,203] Finally, combining benefits and harms as a function and optimizing the function to balance them at the individual level, which may inevitably require weighting of the outcomes[204], would be very relevant for clinical decision-making.

References

1. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *Journal of Clinical Epidemiology*. 2015;68:112-121. doi:10.1016/j.jclinepi.2014.11.010

2. Lassmann H. Multiple sclerosis pathology. *Cold Spring Harbor Perspectives in Medicine*. 2018;8(3):a028936. doi:10.1101/cshperspect.a028936

3. Oh J, Vidal-Jordana A, Montalban X. Multiple sclerosis: clinical aspects. *Current Opinion in Neurology*. 2018;31(6):752-759. doi:10.1097/WCO.0000000000000622

4. 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 VI, Leutmezer F, Limmroth V, Lünemann JD, Lutterotti A, Meuth SG, Meyding-Lamadé U, Platten M, Rieckmann P, Schmidt S, Tumani H, Weber F, Weber MS, Zettl UK, Ziemssen T, Zipp F. Multiple Sclerosis Therapy Consensus Group (MSTCG): position statement on disease-modifying therapies for multiple sclerosis (white paper). *Therapeutic Advances in Neurological Disorders*. 2021;14. doi:10.1177/17562864211039648

5. Kingwell E, Marriott JJ, Jetté N, Pringsheim T, Makhani N, Morrow SA, Fisk JD, Evans C, Béland SG, Kulaga S, Dykeman J, Wolfson C, Koch MW, Marrie RA. Incidence and prevalence of multiple sclerosis in Europe: a systematic review. *BMC Neurology*. 2013;13:128. doi:10.1186/1471-2377-13-128

6. Klein SL, Flanagan KL. Sex differences in immune responses. *Nature Reviews Immunology*. 2016;16(10):626-638. doi:10.1038/nri.2016.90

7. Attfield KE, Jensen LT, Kaufmann M, Friese MA, Fugger L. The immunology of multiple sclerosis. *Nature Reviews Immunology*. 2022:1-17. doi:10.1038/s41577-022-00718-z

8. Hempel S, Fu N, Estrada E, Chen A, Miake-Lye I, Beroes J, Miles J, Shanman R, Shekelle P. *Modifiable risk factors in the progression of multiple sclerosis: a systematic review of the epidemiology and treatment*. 2015. *VA Evidence-based Synthesis Program Reports*.

9. Wallin MT, Culpepper WJ, Nichols E, Bhutta ZA, Gebrehiwot TT, Hay SI, Khalil IA, Krohn KJ, Liang X, Naghavi M, Mokdad AH, Nixon MR, Reiner RC, Sartorius B, Smith M, Topor-Madry R, Werdecker A, Vos T, Feigin VL, Murray CJL. Global, regional, and national burden of multiple sclerosis 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurology*. 2019;18(3):269-285. doi:10.1016/S1474-4422(18)30443-5

10. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, Robertson N, La Rocca N, Uitdehaag B, van der Mei I, Wallin M, Helme A, Angood Napier C, Rijke N, Baneke P. Rising prevalence of multiple sclerosis worldwide: insights from the atlas of MS, third edition. *Multiple Sclerosis*. 2020;26(14):1816-1821. doi:10.1177/1352458520970841

11. Baecher-Allan C, Kaskow BJ, Weiner HL. Multiple sclerosis: mechanisms and immunotherapy. *Neuron*. 2018;97(4):742-768. doi:10.1016/j.neuron.2018.01.021

12. Jacobs BM, Noyce AJ, Bestwick J, Belete D, Giovannoni G, Dobson R. Gene-environment interactions in multiple sclerosis: a UK biobank study. *Neurology - Neuroimmunology Neuroinflammation*. 2021;8(4)doi:10.1212/NXI.0000000000001007

13. Alfredsson L, Olsson T. Lifestyle and environmental factors in multiple sclerosis. *Cold Spring Harbor Perspectives in Medicine*. 2019;9(4):a028944. doi:10.1101/cshperspect.a028944

14. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC, Johnson KP, Sibley WA, Silberberg DH, Tourtellotte WW. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Annals of Neurology*. 1983;13(3):227-231. doi:10.1002/ana.410130302

15. Polman CH, Reingold SC, Edan G, Filippi M, Hartung H-P, Kappos L, Lublin FD, Metz LM, McFarland HF, O'Connor PW, Sandberg-Wollheim M, Thompson AJ, Weinshenker BG, Wolinsky JS. Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". *Annals of Neurology*. 2005;58(6):840-846. doi:10.1002/ana.20703

16. Polman CH, Reingold SC, Banwell B, Clanet M, Cohen JA, Filippi M, Fujihara K, Havrdova E, Hutchinson M, Kappos L, Lublin FD, Montalban X, O'Connor P, Sandberg-Wollheim M, Thompson AJ,

Waubant E, Weinshenker B, Wolinsky JS. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Annals of Neurology*. 2011;69(2):292-302. doi:10.1002/ana.22366

17. McDonald WI, Compston A, Edan G, Goodkin D, Hartung HP, Lublin FD, McFarland HF, Paty DW, Polman CH, Reingold SC, Sandberg-Wollheim M, Sibley W, Thompson A, van den Noort S, Weinshenker BY, Wolinsky JS. Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Annals of Neurology*. 2001;50(1):121-127. doi:10.1002/ana.1032

18. Thompson AJ, Banwell BL, Barkhof F, Carroll WM, Coetzee T, Comi G, Correale J, Fazekas F, Filippi M, Freedman MS, Fujihara K, Galetta SL, Hartung HP, Kappos L, Lublin FD, Marrie RA, Miller AE, Miller DH, Montalban X, Mowry EM, Sorensen PS, Tintoré M, Traboulsee AL, Trojano M, Uitdehaag BMJ, Vukusic S, Waubant E, Weinshenker BG, Reingold SC, Cohen JA. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurology*. 2018;17(2):162-173. doi:10.1016/S1474-4422(17)30470-2

19. Rovira À, Wattjes MP, Tintoré M, Tur C, Yousry TA, Sormani MP, De Stefano N, Filippi M, Auger C, Rocca MA, Barkhof F, Fazekas F, Kappos L, Polman C, Miller D, Montalban X. MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process: evidence-based guidelines. *Nature Reviews: Neurology*. 2015;11(8):471-482. doi:10.1038/nrneurol.2015.106

20. Derfuss T. Personalized medicine in multiple sclerosis: hope or reality? *BMC Medicine*. 2012;10:116. doi:10.1186/1741-7015-10-116

21. Ziemssen T, Akgün K, Brück W. Molecular biomarkers in multiple sclerosis. *Journal of Neuroinflammation*. 2019;16(1):272. doi:10.1186/s12974-019-1674-2

22. Brown FS, Glasmacher SA, Kearns PKA, MacDougall N, Hunt D, Connick P, Chandran S. Systematic review of prediction models in relapsing remitting multiple sclerosis. *PloS One*. 2020;15(5):e0233575. doi:10.1371/journal.pone.0233575

23. Havas J, Leray E, Rollot F, Casey R, Michel L, Lejeune F, Wiertlewski S, Laplaud D, Foucher Y. Predictive medicine in multiple sclerosis: a systematic review. *Multiple Sclerosis and Related Disorders*. 2020;40:101928. doi:10.1016/j.msard.2020.101928

24. Travers BS, Tsang BKT, Barton JL. Multiple sclerosis: diagnosis, disease-modifying therapy and prognosis. *Australian journal of general practice*. 2022;51(4):199-206. doi:10.31128/AJGP-07-21-6103

25. Iacobaeus E, Arrambide G, Amato MP, Derfuss T, Vukusic S, Hemmer B, Tintore M, Brundin L. Aggressive multiple sclerosis (1): towards a definition of the phenotype. *Multiple Sclerosis Journal*. 2020;26(9):1031-1044. doi:10.1177/1352458520925369

26. Diaz C, Zarco LA, Rivera DM. Highly active multiple sclerosis: an update. *Multiple Sclerosis and Related Disorders*. 2019;30:215-224. doi:10.1016/j.msard.2019.01.039

27. Bayas A, Berthele A, Hemmer B, Warnke C, Wildemann B. Controversy on the treatment of multiple sclerosis and related disorders: positional statement of the expert panel in charge of the 2021 DGN Guideline on diagnosis and treatment of multiple sclerosis, neuromyelitis optica spectrum diseases and MOG-IgG-associated disorders. *Neurological research and practice*. 2021;3:45. doi:10.1186/s42466-021-00139-8

28. EPMS. MS treatments. Updated February 15, 2022. Accessed June 9, 2022, https://emsp.org/about-ms/ms-treatments/

29. Lucchetta RC, Oliveira ML, Bonetti AF, Fernandez-Llimos F, Wiens A. Outcome measures for disease-modifying therapies in relapsing multiple sclerosis randomized clinical trials: a scoping review protocol. *JBI evidence synthesis*. 2020;18(8):1781–1787. doi:10.11124/JBISRIR-D-19-00178

30. Lucchetta RC, Tonin FS, Borba HHL, Leonart LP, Ferreira VL, Bonetti AF, Riveros BS, Becker J, Pontarolo R, Fernandez-Llimós F, Wiens A. Disease-modifying therapies for relapsing–remitting multiple sclerosis: a network meta-analysis. *CNS Drugs*. 2018;32(9):813-826. doi:10.1007/s40263-018-0541-5

31. Oliveira ML, Lucchetta RC, Bonetti AdF, Fernandez-Llimós F, Becker J, Gonçalves MVM, Tauil CB, Pontarolo R, Wiens A. Efficacy outcomes reported in trials of multiple sclerosis: A systematic scoping review. *Multiple Sclerosis and Related Disorders*. 2020;45:102435. doi:10.1016/j.msard.2020.102435

32. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). *Neurology*. 1983;33(11):1444-1452. doi:10.1212/WNL.33.11.1444

33. van Munster CEP, Uitdehaag BMJ. Outcome measures in clinical trials for multiple sclerosis. *CNS Drugs*. 2017;31(3):217-236. doi:10.1007/s40263-017-0412-5

34. Barkhof F. The clinico-radiological paradox in multiple sclerosis revisited. *Current Opinion in Neurology*. 2002;15(3):239-245. doi:10.1097/00019052-200206000-00003

35. Sormani MP, Bonzano L, Roccatagliata L, Cutter GR, Mancardi GL, Bruzzi P. Magnetic resonance imaging as a potential surrogate for relapses in multiple sclerosis: a meta-analytic approach. *Annals of Neurology*. 2009;65(3):268-275. doi:10.1002/ana.21606

36. Sormani MP, Bruzzi P. MRI lesions as a surrogate for relapses in multiple sclerosis: a meta-analysis of randomised trials. *Lancet Neurology*. 2013;12(7):669-676. doi:10.1016/S1474-4422(13)70103-0

37. Inojosa H, Schriefer D, Ziemssen T. Clinical outcome measures in multiple sclerosis: a review. *Autoimmunity reviews*. 2020;19(5):102512. doi:10.1016/j.autrev.2020.102512

38. Day GS, Rae-Grant A, Armstrong MJ, Pringsheim T, Cofield SS, Marrie RA. Identifying priority outcomes that influence selection of disease-modifying therapies in MS. *Neurology Clinical practice*. 2018;8(3):179-185. doi:10.1212/CPJ.0000000000000449

39. Chen C, Zhang E, Zhu C, Wei R, Ma L, Dong X, Li R, Sun F, Zhou Y, Cui Y, Liu Z. Comparative efficacy and safety of disease-modifying therapies in patients with relapsing multiple sclerosis: A systematic review and network meta-analysis. *Journal of the American Pharmacists Association*. 2023;63:8-22. doi:10.1016/j.japh.2022.07.009

40. Rae-Grant A, Day GS, Marrie RA, Rabinstein A, Cree BAC, Gronseth GS, Haboubi M, Halper J, Hosey JP, Jones DE, Lisak R, Pelletier D, Potrebic S, Sitcov C, Sommers R, Stachowiak J, Getchius TSD, Merillat SA, Pringsheim T. Practice guideline recommendations summary: disease-modifying therapies for adults with multiple sclerosis: report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. *Neurology*. 2018;90(17):777-788. doi:10.1212/WNL.0000000000005347

41. Montalban X, Gold R, Thompson AJ, Otero-Romero S, Amato MP, Chandraratna D, Clanet M, Comi G, Derfuss T, Fazekas F, Hartung HP, Havrdova E, Hemmer B, Kappos L, Liblau R, Lubetzki C, Marcus E, Miller DH, Olsson T, Pilling S, Selmaj K, Siva A, Sorensen PS, Sormani MP, Thalheim C, Wiendl H, Zipp F. ECTRIMS/EAN Guideline on the pharmacological treatment of people with multiple sclerosis. *Multiple Sclerosis Journal*. 2018;24(2):96-120. doi:10.1177/1352458517751049

42. Stankiewicz JM, Weiner HL. An argument for broad use of high efficacy treatments in early multiple sclerosis. *Neurology(R) neuroimmunology & neuroinflammation*. 2020;7(1):e636. doi:10.1212/NXI.0000000000000636

43. Weideman AM, Tapia-Maltos MA, Johnson K, Greenwood M, Bielekova B. Meta-analysis of the age-dependent efficacy of multiple sclerosis treatments. *Frontiers in Neurology*. 2017;8:577. doi:10.3389/fneur.2017.00577

44. Vermersch P, Berger T, Gold R, Lukas C, Rovira A, Meesen B, Chard D, Comabella M, Palace J, Trojano M. The clinical perspective: How to personalise treatment in MS and how may biomarkers including imaging contribute to this? *Multiple Sclerosis*. 2016;22(2S):18-33. doi:10.1177/1352458516650739

45. Lipkovich I, Dmitrienko A, D'Agostino Sr. RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*. 2017;36(1):136-196. doi:10.1002/sim.7064

46. Signori A, Schiavetti I, Gallo F, Sormani MP. Subgroups of multiple sclerosis patients with larger treatment benefits: a meta-analysis of randomized trials. *European Journal of Neurology*. 2015;22(6):960-966. doi:10.1111/ene.12690

47. Winkelmann A, Loebermann M, Reisinger EC, Hartung H-P, Zettl UK. Disease-modifying therapies and infectious risks in multiple sclerosis. *Nature Reviews: Neurology*. 2016;12(4):217-233. doi:10.1038/nrneurol.2016.21

48. Crommelin DJA, Broich K, Holloway C, Meesen B, Lizrova Preiningerova J, Prugnaud J-L, Silva-Lima B. The regulator's perspective: How should new therapies and follow-on products for MS be

clinically evaluated in the future? *Multiple Sclerosis*. 2016;22(2S):47-59. doi:10.1177/1352458516650744

49. Gafson A, Craner MJ, Matthews PM. Personalised medicine for multiple sclerosis care. *Multiple Sclerosis*. 2017;23(3):362-369. doi:10.1177/1352458516672017

50. Li H, Hu F, Zhang Y, Li K. Comparative efficacy and acceptability of disease-modifying therapies in patients with relapsing–remitting multiple sclerosis: a systematic review and network meta-analysis. *Journal of Neurology*. 2019;doi:10.1007/s00415-019-09395-w

51. Farber RS, Sand IK. Optimizing the initial choice and timing of therapy in relapsing-remitting multiple sclerosis. *Therapeutic Advances in Neurological Disorders*. 2015;8(5):212-232. doi:10.1177/1756285615598910

52. Gilenya [prescribing information]. East Hanover, New Jersey: Novartis Pharmaceuticals Corporation; 2018.

53. Gilenya [summary of product characteristics]. Dublin, Ireland: Novartis Europharm Limited; 2020.

54. 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. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *New England Journal of Medicine*. 2010;362(5):387-401. doi:10.1056/NEJMoa0909494

55. Calabresi PA, Radue E-W, Goodin D, Jeffery D, Rammohan KW, Reder AT, Vollmer T, Agius MA, Kappos L, Stites T, Li B, Cappiello L, von Rosenstiel P, Lublin FD. Safety and efficacy of fingolimod in patients with relapsing-remitting multiple sclerosis (FREEDOMS II): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Neurology*. 2014;13(6):545-556. doi:10.1016/S1474-4422(14)70049-3

56. Cohen JA, Barkhof F, Comi G, Hartung H-P, Khatri BO, Montalban X, Pelletier J, Capra R, Gallo P, Izquierdo G, Tiel-Wilck K, de Vera A, Jin J, Stites T, Wu S, Aradhye S, Kappos L. Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis. *New England Journal of Medicine*. 2010;362(5):402-415. doi:10.1056/NEJMoa0907839

57. Jalkh G, Abi Nahed R, Macaron G, Rensel M. Safety of newer disease modifying therapies in multiple sclerosis. *Vaccines*. 2020;9(1):12. doi:10.3390/vaccines9010012

58. Mantia LL, Tramacere I, Firwana B, Pacchetti I, Palumbo R, Filippini G. Fingolimod for relapsing-remitting multiple sclerosis. *Cochrane Database of Systematic Reviews*. 2016;(4)doi:10.1002/14651858.CD009371.pub2

59. Zhao Z, Ma C-L, Gu Z-C, Dong Y, Lv Y, Zhong M-K. Incidence and risk of infection associated with fingolimod in patients with multiple sclerosis: a systematic review and meta-analysis of 8,448 patients from 12 randomized controlled trials. *Frontiers in Immunology*. 2021;12:611711. doi:10.3389/fimmu.2021.611711

60. Askari M, Mirmosayyeb O, Ghaffary EM, Ghoshouni H, Shaygannejad V, Ghajarzadeh M. Incidence of cancer in patients with multiple sclerosis (MS) who were treated with fingolimod: a systematic review and meta-analysis. *Multiple Sclerosis and Related Disorders*. 2022;59:103680. doi:10.1016/j.msard.2022.103680

61. Alping P, Askling J, Burman J, Fink K, Fogdell-Hahn A, Gunnarsson M, Hillert J, Langer-Gould A, Lycke J, Nilsson P, Salzer J, Svenningsson A, Vrethem M, Olsson T, Piehl F, Frisell T. Cancer risk for fingolimod, natalizumab, and rituximab in multiple sclerosis patients. *Annals of Neurology*. 2020;87(5):688-699. doi:10.1002/ana.25701

62. La Mantia L, Benedetti MD, Sant M, d'Arma A, Di Tella S, Lillini R, Mendozzi L, Marangi A, Turatti M, Caputo D, Rovaris M. Cancer risk for multiple sclerosis patients treated with azathioprine and disease-modifying therapies: an Italian observational study. *Neurological Sciences*. 2021;42(12):5157-5163. doi:10.1007/s10072-021-05216-z

63. Stamatellos V-P, Siafis S, Papazisis G. Disease-modifying agents for multiple sclerosis and the risk for reporting cancer: a disproportionality analysis using the US Food and Drug Administration adverse event reporting system database. *British Journal of Clinical Pharmacology*. 2021;87(12):4769-4779. doi:10.1111/bcp.14916

64. Devonshire V, Havrdova E, Radue EW, O'Connor P, Zhang-Auberson L, Agoropoulou C, Häring DA, Francis G, Kappos L. Relapse and disability outcomes in patients with multiple sclerosis treated

with fingolimod: subgroup analyses of the double-blind, randomised, placebo-controlled FREEDOMS study. *Lancet Neurology*. 2012;11(5):420-428. doi:10.1016/S1474-4422(12)70056-X

65. Derfuss T, Ontaneda D, Nicholas J, Meng X, Hawker K. Relapse rates in patients with multiple sclerosis treated with fingolimod: subgroup analyses of pooled data from three phase 3 trials. *Multiple Sclerosis and Related Disorders*. 2016;8:124-130. doi:10.1016/j.msard.2016.05.015

66. Hayden JA, Côté P, Steenstra IA, Bombardier C. Identifying phases of investigation helps planning, appraising, and applying the results of explanatory prognosis studies. *Journal of Clinical Epidemiology*. 2008;61(6):552-560. doi:10.1016/j.jclinepi.2007.08.005

67. van der Leeuw J, Ridker PM, van der Graaf Y, Visseren FLJ. Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. *European Heart Journal*. 2014;35(13):837-843. doi:10.1093/eurheartj/ehu004

68. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ*. 2009;338:b375. doi:10.1136/bmj.b375

69. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. doi:10.1097/EDE.0b013e3181c30fb2

70. Gerds TA, Cai T, Schumacher M. The Performance of Risk Prediction Models. *Biometrical Journal*. 2008;50(4):457-479. doi:10.1002/bimj.200810443

71. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, Malats N, Briggs A, Schroter S, Altman DG, Hemingway H, Group ftP. Prognosis research strategy (PROGRESS) 2: prognostic factor research. *PLoS Medicine*. 2013;10(2):e1001380. doi:10.1371/journal.pmed.1001380

72. Steyerberg EW, Moons KGM, Windt DAvd, Hayden JA, Perel P, Schroter S, Riley RD, Hemingway H, Altman DG, Group ftP. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine*. 2013;10(2):e1001381. doi:10.1371/journal.pmed.1001381

73. Sormani MP, Rio J, Tintorè M, Signori A, Li D, Cornelisse P, Stubinski B, Stromillo ML, Montalban X, De Stefano N. Scoring treatment response in patients with relapsing multiple sclerosis. *Multiple Sclerosis*. 2013;19(5):605-612. doi:10.1177/1352458512460605

74. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JaH. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis*. 2013;11(s1):129-141. doi:10.1111/jth.12262

75. VanderWeele TJ, Luedtke AR, van der Laan MJ, Kessler RC. Selecting optimal subgroups for treatment using many covariates. *Epidemiology*. 2019;30(3):334-341. doi:10.1097/EDE.0000000000000991

76. Hemingway H, Riley RD, Altman DG. Ten steps towards improving prognosis research. *BMJ*. 2009;339:b4184. doi:10.1136/bmj.b4184

77. Shmueli G. To explain or to predict? *Statistical Science*. 2010;25(3):289-310. doi:10.1214/10-STS330

78. Kent DM, van Klaveren D, Paulus JK, D'Agostino R, Goodman S, Hayward R, Ioannidis JPA, Patrick-Lake B, Morton S, Pencina M, Raman G, Ross JS, Selker HP, Varadhan R, Vickers A, Wong JB, Steyerberg EW. The predictive approaches to treatment effect heterogeneity (PATH) statement: explanation and elaboration. *Annals of Internal Medicine*. 2020;172(1):W1-W25. doi:10.7326/M18-3668

79. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*. 2011;30(24):2867-2880. doi:10.1002/sim.4322

80. Cai T, Tian L, Wong PH, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*. 2011;12(2):270-282. doi:10.1093/biostatistics/kxq060

81. Janes H, Pepe MS, McShane LM, Sargent DJ, Heagerty PJ. The fundamental difficulty with evaluating the accuracy of biomarkers for guiding treatment. *Journal of the National Cancer Institute*. 2015;107(8):djv157. doi:10.1093/jnci/djv157

82. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine*. 2000;19(4):453-473. doi:10.1002/(SICI)1097-0258(20000229)19:4<453::AID-SIM350>3.0.CO;2-5

83. Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. *Journal of Clinical Epidemiology*. 2008;61(11):1085-1094. doi:10.1016/j.jclinepi.2008.04.008

84. Steyerberg EW. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. 2nd ed. Statistics for Biology and Health. Springer Nature Switzerland AG; 2019.

85. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

86. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. *BMJ*. 2009;338:b604. doi:10.1136/bmj.b604

87. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology*. 2013;66(8):818-825. doi:10.1016/j.jclinepi.2013.02.009

88. Simon R, Altman D. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*. 1994;69:979-985. doi:10.1038/bjc.1994.192

89. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell Jr FE, Moons KGM, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*. 2019;38(7):1276-1296. doi:10.1002/sim.7992

90. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*. 2001;54(8):774-781. doi:10.1016/S0895-4356(01)00341-9

91. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of Internal Medicine*. 2019;170(1):W1. doi:10.7326/M18-1377

92. Groenwold RHH, Moons KGM, Pajouheshnia R, Altman DG, Collins GS, Debray TPA, Reitsma JB, Riley RD, Peelen LM. Explicit inclusion of treatment in prognostic modeling was recommended in observational and randomized settings. *Journal of Clinical Epidemiology*. 2016;78:90-100. doi:10.1016/j.jclinepi.2016.03.017

93. Janes H, Brown MD, Huang Y, Pepe MS. An approach to evaluating and comparing biomarkers for patient treatment selection. *The international journal of biostatistics*. 2014;10(1):99-121. doi:10.1515/ijb-2012-0052

94. Kent DM, Steyerberg E, Klaveren Dv. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*. 2018;363doi:10.1136/bmj.k4245

95. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *International Journal of Epidemiology*. 2016;45(6):2184-2193. doi:10.1093/ije/dyw125

96. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*. 2000;355(9209):1064-1069. doi:10.1016/S0140-6736(00)02039-0

97. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine — reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*. 2007;357(21):2189-2194. doi:10.1056/NEJMsr077003

98. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*. 2010;340:c117. doi:10.1136/bmj.c117

99. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, Bala MM, Bassler D, Mertz D, Diaz-Granados N, Vandvik PO, Malaga G, Srinathan SK, Dahm P, Johnston BC, Alonso-Coello P, Hassouneh B, Walter SD, Heels-Ansdell D, Bhatnagar N, Altman DG, Guyatt GH. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ*. 2012;344:e1553. doi:10.1136/bmj.e1553

100. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*. 2011;30:2601-2621. doi:10.1002/sim.4289

101. Hingorani AD, Windt DAvd, Riley RD, Abrams K, Moons KGM, Steyerberg EW, Schroter S, Sauerbrei W, Altman DG, Hemingway H. Prognosis research strategy (PROGRESS) 4: stratified medicine research. *BMJ*. 2013;346:e5793. doi:10.1136/bmj.e5793

102. Hoogland J, IntHout J, Belias M, Rovers MM, Riley RD, E. Harrell Jr F, Moons KGM, Debray TPA, Reitsma JB. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in Medicine*. 2021;40(26):5961-5981. doi:10.1002/sim.9154

103. Rekkas A, Paulus JK, Raman G, Wong JB, Steyerberg EW, Rijnbeek PR, Kent DM, van Klaveren D. Predictive approaches to heterogeneous treatment effects: a scoping review. *BMC Medical Research Methodology*. 2020;20:264. doi:10.1186/s12874-020-01145-1

104. Janes H, Pepe MS, Bossuyt PM, Barlow WE. Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*. 2011;154(4):253-259. doi:10.7326/0003-4819-154-4-201102150-00006

105. Seibold H, Zeileis A, Hothorn T. Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Statistical Methods in Medical Research*. 2018;27(10):3104-3125. doi:10.1177/0962280217693034

106. Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, Stallard N, Posch M. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *Journal of Biopharmaceutical Statistics*. 2016;26(1):99-119. doi:10.1080/10543406.2015.1092034

107. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*. 2016;113(27):7353-7360. doi:10.1073/pnas.1510489113

108. Mansmann U, Rieger A, Strahwald B, Crispin A. Risk calculators—methods, development, implementation, and validation. *International Journal of Colorectal Disease*. 2016;31:1111-1116. doi:10.1007/s00384-016-2589-3

109. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996;15(4):361-387. doi:10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

110. Justice AC. Assessing the generalizability of prognostic information. *Annals of Internal Medicine*. 1999;130(6):515-524. doi:10.7326/0003-4819-130-6-199903160-00016

111. Boulesteix AL, Janitza S, Hornung R, Probst P, Busen H, Hapfelmeier A. Making complex prediction rules applicable for readers: current practice in random forest literature and recommendations. *Biometrical Journal*. 2019;61(5):1314-1328. doi:10.1002/bimj.201700243

112. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605. doi:10.1136/bmj.b605

113. Hoogland J, Efthimiou O, Nguyen TL, Debray TPA. Evaluating individualized treatment effect predictions: a new perspective on discrimination and calibration assessment. *arXiv*. 2022. Preprint posted online September 13, 2022. doi:10.48550/arXiv.2209.06101 www.arxiv.org/abs/2209.06101

114. Adams ST, Leveson SH. Clinical prediction rules. *BMJ*. 2012;344:d8312. doi:10.1136/bmj.d8312

115. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *European Heart Journal*. 2014;35(29):1925-1931. doi:10.1093/eurheartj/ehu207

116. Gneiting T, Raftery AE. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*. 2007;102(477):359-378. doi:10.1198/016214506000001437

117. Vickers AJ, Cronin AM. Everything you always wanted to know about evaluating prediction models (but were too afraid to ask). *Urology*. 2010;76(6):1298-1301. doi:10.1016/j.urology.2010.06.019

118. Vickers AJ, Calster BV, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*. 2016;352:i6. doi:10.1136/bmj.i6

119. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and prognostic research*. 2019;3(1):18. doi:10.1186/s41512-019-0064-7

120. Maas CCHM, Kent DM, Hughes MC, Dekker R, Lingsma HF, Klaveren Dv. Performance metrics for models designed to predict treatment effect. *medRxiv*. 2023. Preprint posted online January 13, 2023. doi:10.1101/2022.06.14.22276387 www.medrxiv.org/content/10.1101/2022.06.14.22276387v3

121. Xia Y, Gustafson P, Sadatsafavi M. Methodological concerns about 'concordance-statistic for benefit' as a measure of discrimination in treatment benefit prediction. *arXiv*. 2022. Preprint posted online December 8, 2022. doi:10.48550/arXiv.2208.13553 www.arxiv.org/abs/2208.13553

122. Chalkou K, Steyerberg E, Egger M, Manca A, Pellegrini F, Salanti G. A two-stage prediction model for heterogeneous effects of treatments. *Statistics in Medicine*. 2021;40(20):4362-4375. doi:10.1002/sim.9034

123. Trojano M, Tintore M, Montalban X, Hillert J, Kalincik T, Iaffaldano P, Spelman T, Sormani MP, Butzkueven H. Treatment decisions in multiple sclerosis — insights from real-world observational studies. *Nature Reviews: Neurology*. 2017;13(2):105-118. doi:10.1038/nrneurol.2016.188

124. Rotstein D, Montalban X. Reaching an evidence-based prognosis for personalized treatment of multiple sclerosis. *Nature Reviews Neurology*. 2019;15(5):287-300. doi:10.1038/s41582-019-0170-8

125. Dennison L, Brown M, Kirby S, Galea I. Do people with multiple sclerosis want to know their prognosis? A UK nationwide study. *PloS One*. 2018;13(2):e0193407. doi:10.1371/journal.pone.0193407

126. Ziemssen T, Stefano ND, Sormani MP, Wijmeersch BV, Wiendl H, Kieseier BC. Optimizing therapy early in multiple sclerosis: an evidence-based view. *Multiple Sclerosis and Related Disorders*. 2015;4(5):460-469. doi:10.1016/j.msard.2015.07.007

127. Hemond CC, Bakshi R. Magnetic Resonance Imaging in Multiple Sclerosis. *Cold Spring Harbor Perspectives in Medicine*. 2018;8(5):a028969. doi:10.1101/cshperspect.a028969

128. Sormani MP. Prognostic factors versus markers of response to treatment versus surrogate endpoints: three different concepts. *Multiple Sclerosis*. 2017;23(3):378-381. doi:10.1177/1352458516676899

129. Reeve K, On BI, Havla J, Burns J, Gosteli-Peter MA, Alabsawi A, Alayash Z, Götschi A, Seibold H, Mansmann U, Held U. Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. *Cochrane Database of Systematic Reviews*. 2023;2023(9)doi:10.1002/14651858.CD013606.pub2

130. Bergamaschi R, Berzuini C, Romani A, Cosi V. Predicting secondary progression in relapsing–remitting multiple sclerosis: a Bayesian analysis. *Journal of the Neurological Sciences*. 2001;189(1):13-21. doi:10.1016/S0022-510X(01)00572-X

131. Manouchehrinia A, Zhu F, Piani-Meier D, Lange M, Silva DG, Carruthers R, Glaser A, Kingwell E, Tremlett H, Hillert J. Predicting risk of secondary progression in multiple sclerosis: a nomogram. *Multiple Sclerosis*. 2019;25(8):1102-1112. doi:10.1177/1352458518783667

132. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, Group† ftP. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*. 2019;170(1):51. doi:10.7326/M18-1376

133. Pellegrini F, Copetti M, Sormani MP, Bovis F, de Moor C, Debray TPA, Kieseier BC. Predicting disability progression in multiple sclerosis: insights from advanced statistical modeling. *Multiple Sclerosis*. 2020;26(14):1828-1836. doi:10.1177/1352458519887343

134. De Brouwer E, Becker T, Moreau Y, Havrdova EK, 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 JL, Solaro C, Boz C, Iuliano G, Buzzard K, Aguera-Morales E, Terzi M, Trivio TC, Spitaleri D, Van Pesch V, Shaygannejad V, Moore F, Oreja-Guevara C, Maimone D, Gouider R, Csepany T, Ramo-Tello C, Peeters L. Longitudinal machine learning modeling of MS patient trajectories improves predictions of disability progression. *Computer Methods and Programs in Biomedicine*. 2021;208:106180. doi:10.1016/j.cmpb.2021.106180

135. De Brouwer E, Becker T, Moreau Y, Havrdova EK, 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 MJ, Van Wijmeersch B, Soysal A, Sanchez-Menoyo JL, Solaro C, Boz C, Iuliano G, Buzzard K, Aguera-Morales E, Terzi M, Trivio TC, Spitaleri D, Van Pesch V, Shaygannejad V, Moore F, Oreja-Guevara C, Maimone D, Gouider R, Csepany T, Ramo-Tello C, Peeters L. 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]. *Computer Methods and Programs in Biomedicine*. 2022;213:106479. doi:10.1016/j.cmpb.2021.106479

136. Chalkou K, Steyerberg E, Bossuyt P, Subramaniam S, Benkert P, Kuhle J, Disanto G, Kappos L, Zecca C, Egger M, Salanti G. Development, validation and clinical usefulness of a prognostic model for relapse in relapsing-remitting multiple sclerosis. *Diagnostic and prognostic research*. 2021;5(1):17. doi:10.1186/s41512-021-00106-6

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

138. Bovis F, Carmisciano L, Signori A, Pardini M, Steinerman JR, Li T, Tansy AP, Sormani MP. Defining responders to therapies by a statistical modeling approach applied to randomized clinical trial data. *BMC Medicine*. 2019;17(1)doi:10.1186/s12916-019-1345-2

139. Pellegrini F, Copetti M, Bovis F, Cheng D, Hyde R, de Moor C, Kieseier BC, Sormani MP. A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. *Multiple Sclerosis*. 2020;26(9):1064-1073. doi:10.1177/1352458519849513

140. Stühler E, Braune S, Lionetto F, Heer Y, Jules E, Westermann C, Bergmann A, van Hövell P, NeuroTransData Study G. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Medical Research Methodology*. 2020;20(1):24. doi:10.1186/s12874-020-0906-6

141. Kalincik T, Kister I, Bacon TE, Malpas CB, Sharmin S, Horakova D, Kubala-Havrdova E, Patti F, Izquierdo G, Eichau S, Ozakbas S, Onofrj M, Lugaresi A, Prat A, Girard M, Duquette P, Grammond P, Sola P, Ferraro D, Alroughani R, Terzi M, Boz C, Grand'Maison F, Bergamaschi R, Gerlach O, Sa MJ, Kappos L, Cartechini E, Lechner-Scott J, van Pesch V, Shaygannejad V, Granella F, Spitaleri D, Iuliano G, Maimone D, Prevost J, Soysal A, Turkoglu R, Ampapa R, Butzkueven H, Cutter G. Multiple sclerosis severity score (MSSS) improves the accuracy of individualized prediction in MS. *Multiple Sclerosis Journal*. 2022;28(11):1752-1761. doi:10.1177/13524585221084577

142. Brouwer ED, Becker T, Werthen-Brabants L, Dewulf P, Iliadis D, Dekeyser C, Laureys G, Wijmeersch BV, Popescu V, Dhaene T, Deschrijver D, Waegeman W, Baets BD, Stock M, Horakova D, Patti F, Izquierdo G, Eichau S, Girard M, Prat A, Lugaresi A, Grammond P, Kalincik T, Alroughani R, Grand'Maison F, Skibina O, Terzi M, Lechner-Scott J, Gerlach O, Khoury SJ, Cartechini E, Pesch VV, Sa MJ, Weinstock-Guttman B, Blanco Y, Ampapa R, Spitaleri D, Solaro C, Maimone D, Soysal A, Iuliano G, Gouider R, Castillo-Triviño T, Sanchez-Menoyo JL, Laureys G, Walt Avd, Oh J, Aguera-Morales E, Altintas A, Al-Asmi A, Gans Kd, Fragoso Y, Csepany T, Hodgkinson S, Deri N, Al-Harbi T, Taylor B, Gray O, Lalive P, Rozsa C, McGuigan C, Kermode A, Sempere AP, Mihaela S, Simo M, Hardy T, Decoo D, Hughes S, Grigoriadis N, Sas A, Vella N, Moreau Y, Peeters L. Machine-learning-based prediction of disability progression in multiple sclerosis: an observational, international, multi-center study. *medRxiv*. 2022. Preprint posted online September 11, 2022. doi:10.1101/2022.09.08.22279617 www.medrxiv.org/content/10.1101/2022.09.08.22279617v1

143. Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*. 2009;28(26):3294-3315. doi:10.1002/sim.3720

144. Steyerberg EW, Claggett B. Towards personalized therapy for multiple sclerosis: limitations of observational data. *Brain*. 2018;141(5):e38-e38. doi:10.1093/brain/awy055

145. Kalincik T. Reply: Towards personalized therapy for multiple sclerosis: limitations of observational data. *Brain*. 2018;141(5):e39-e39. doi:10.1093/brain/awy056

146. Braune S, Stuehler E, Heer Y, van Hoevell P, Bergmann A, NSG. PHREND®—a real-world data-driven tool supporting clinical decisions to optimize treatment in relapsing-remitting multiple sclerosis. Original Research. *Frontiers in digital health*. 2022;4:856829. doi:10.3389/fdgth.2022.856829

147. OFSEP. EXTVAL-PHREND Validation externe des prédictions de réponse individualisée au traitement chez les patients atteints de sclérose en plaques récurrente-rémittente. Accessed February 17, 2023, https://www.ofsep.org/en/studies/extval-phrend

148. Vukusic S, Casey R, Rollot F, Brochet B, Pelletier J, Laplaud D-A, De Sèze J, Cotton F, Moreau T, Stankoff B, Fontaine B, Guillemin F, Debouverie M, Clanet M. Observatoire Français de la Sclérose en Plaques (OFSEP): A unique multimodal nationwide MS registry in France. *Multiple Sclerosis*. 2018;26(1):118-122. doi:10.1177/1352458518815602

149. DIFUTURE. Data integration for future medicine. Accessed December 14, 2022, https://difuture.de/

150. DIFUTURE. Approach. Accessed December 14, 2022, https://difuture.de/our-approach/

151. Hapfelmeier A, On BI, Mühlau M, Kirschke JS, Berthele A, Gasperi C, Mansmann U, Wuschek A, Bussas M, Boeker M, Bayas A, Senel M, Havla J, Kowarik MC, Kuhn K, Gatz I, Spengler H, Wiestler B, Grundl L, Sepp D, Hemmer B. Retrospective cohort study to devise a treatment decision score predicting adverse 24-month radiological activity in early multiple sclerosis. *Therapeutic Advances in Neurological Disorders*. 2023;16:1-25. doi:10.1177/17562864231161892

152. DRKS. Deutsches Register Klinischer Studien. Accessed February 27, 2023, https://drks.de/search/de

153. CSDR. Clinical Study Data Request. Accessed February 27, 2023, www.clinicalstudydatarequest.com

154. EuroQol Group. EuroQol--a new facility for the measurement of health-related quality of life. *Health Policy*. 1990;16(3):199-208. doi:10.1016/0168-8510(90)90421-9

155. Nicholas R, Straube S, Schmidli H, Pfeiffer S, Friede T. Time-patterns of annualized relapse rates in randomized placebo-controlled clinical trials in relapsing multiple sclerosis: a systematic review and meta-analysis. *Multiple Sclerosis*. 2012;18(9):1290-1296. doi:10.1177/1352458511435715

156. Siri P, Henninger E, Sormani MP. A parametric model fitting time to first event for overdispersed data: application to time to relapse in multiple sclerosis. *Lifetime Data Analysis*. 2012;18(2):139-156. doi:10.1007/s10985-011-9207-z

157. George B, Seals S, Aban I. Survival analysis and regression models. *Journal of Nuclear Cardiology*. 2014;21(4):686-694. doi:10.1007/s12350-014-9908-2

158. Sormani M, Signori A, Siri P, De Stefano N. Time to first relapse as an endpoint in multiple sclerosis clinical trials. *Multiple Sclerosis*. 2013;19(4):466-474. doi:10.1177/1352458512457841

159. Wang YC, Meyerson L, Tang YQ, Qian N. Statistical methods for the analysis of relapse data in MS clinical trials. *Journal of the Neurological Sciences*. 2009;285(1):206-211. doi:10.1016/j.jns.2009.07.017

160. Hothorn T, Möst L, Bühlmann P. Most likely transformations. *Scandinavian journal of statistics, theory and applications*. 2018;45(1):110-134. doi:https://doi.org/10.1111/sjos.12291

161. Hothorn T, Kneib T, Bühlmann P. Conditional transformation models. *Journal of the Royal Statistical Society Series B, Statistical methodology*. 2014;76(1):3-27.

162. Korepanova N, Seibold H, Steffen V, Hothorn T. Survival forests under test: impact of the proportional hazards assumption on prognostic and predictive forests for amyotrophic lateral sclerosis survival. *Statistical Methods in Medical Research*. 2020;29(5):1403-1419. doi:10.1177/0962280219862586

163. Seibold H, Zeileis A, Hothorn T. Model-based recursive partitioning for subgroup analyses. *The international journal of biostatistics*. 2016;12(1):45-63. doi:10.1515/ijb-2015-0032

164. Hothorn T, Zeileis A. Predictive distribution modeling using transformation forests. *Journal of computational and graphical statistics*. 2021;30(4):1181-1196. doi:10.1080/10618600.2021.1872581

165. Mentch L, Zhou S. Randomization as regularization: a degrees of freedom explanation for random forest success. *Journal of machine learning research : JMLR*. 2020;21(171):1-36. doi:10.48550/ARXIV.1911.00190

166. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*. 2009;14(4):323-348. doi:10.1037/a0016973

167. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: with Applications in R*. 2nd ed. Springer US; 2021.

168. Hapfelmeier A. *Analysis of missing data with random forests*. Dissertation. LMU Munich; 2012.

169. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B Statistical Methodology*. 2005;67(2):301-320. doi:10.1111/j.1467-9868.2005.00503.x

170. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Statistics and its interface*. 2009;2(3):369-380. doi:10.4310/sii.2009.v2.n3.a10

171. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B Statistical Methodology*. 2006;68(1):49-67. doi:10.1111/j.1467-9868.2005.00532.x

172. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2011;28(1):112-118. doi:10.1093/bioinformatics/btr597

173. Hapfelmeier A, Hothorn T, Ulm K, Strobl C. A new variable importance measure for random forests with missing data. *Statistics and computing*. 2014;24(1):21-34. doi:10.1007/s11222-012-9349-1

174. Buuren Sv, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *Journal of statistical software*. 2011;45:1-67. doi:10.18637/jss.v045.i03

175. Hoogland J, van Barreveld M, Debray TPA, Reitsma JB, Verstraelen TE, Dijkgraaf MGW, Zwinderman AH. Handling missing predictor values when validating and applying a prediction model to new patients. *Statistics in Medicine*. 2020;39(25):3591-3607. doi:10.1002/sim.8682

176. Wan Y, Datta S, Conklin DJ, Kong M. Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of statistical computation and simulation*. 2015;85(9):1902-1916. doi:10.1080/00949655.2014.907801

177. White IR, Royston P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*. 2009;28(15):1982-1998. doi:10.1002/sim.3618

178. Bansal A, Heagerty PJ. A tutorial on evaluating the time-varying discrimination accuracy of survival models used in dynamic decision making. *Medical Decision Making*. 2018;38(8):904-916. doi:10.1177/0272989X18801312

179. Bansal A, Heagerty PJ. A comparison of landmark methods and time-dependent ROC methods to evaluate the time-varying performance of prognostic markers for survival outcomes. *Diagnostic and prognostic research*. 2019;3(1):14. doi:10.1186/s41512-019-0057-6

180. Blanche P, Kattan MW, Gerds TA. The c-index is not proper for the evaluation of *t*-year predicted risks. *Biostatistics*. 2019;20(2):347-357. doi:10.1093/biostatistics/kxy006

181. Probst P, Wright MN, Boulesteix A-L. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*. 2019;9(3):e1301. doi:10.1002/widm.1301

182. Probst P, Boulesteix A-L, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*. 2019;20:1-32.

183. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*. 2012;50(11):1-23.

184. Gerds TA, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*. 2006;48(6):1029-1040. doi:10.1002/bimj.200610301

185. Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*. 2016;25(4):1692-1706. doi:10.1177/0962280213497434

186. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Medical Informatics and Decision Making*. 2008;8:53. doi:10.1186/1472-6947-8-53

187. van Klaveren D, Vergouwe Y, Farooq V, Serruys PW, Steyerberg EW. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *Journal of Clinical Epidemiology*. 2015;68(11):1366-1374. doi:10.1016/j.jclinepi.2015.02.012

188. Molnar C, Casalicchio G, Bischl B. Interpretable machine learning – a brief history, state-of-the-art and challenges. ECML PKDD 2020 Workshops. Springer, Cham; 2020:417-431.

189. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*. 2016;74:167-176. doi:10.1016/j.jclinepi.2015.12.005

190. Niederberger E, Parnham MJ. The impact of diet and exercise on drug responses. *International Journal of Molecular Sciences*. 2021;22(14):7692. doi:10.3390/ijms22147692

191. Mathiowetz V, Weber K, Kashman N, Volland G. Adult norms for the nine hole peg test of finger dexterity. *The Occupational Therapy Journal of Research*. 1985;5(1):24-38. doi:10.1177/153944928500500102

192. Mandel M, Mercier F, Eckert B, Chin P, Betensky RA. Estimating time to disease progression comparing transition models and survival methods—an analysis of multiple sclerosis data. *Biometrics*. 2013;69(1):225-234. doi:10.1111/biom.12002

193. Mallon A-M, Häring DA, Dahlke F, Aarden P, Afyouni S, Delbarre D, El Emam K, Ganjgahi H, Gardiner S, Kwok CH, West DM, Straiton E, Haemmerle S, Huffman A, Hofmann T, Kelly LJ, Krusche P, Laramee M-C, Lheritier K, Ligozio G, Readie A, Santos L, Nichols TE, Branson J, Holmes C. Advancing data science in drug development through an innovative computational framework for data sharing and statistical analysis. *BMC Medical Research Methodology*. 2021;21(1):250. doi:10.1186/s12874-021-01409-4

194. Wynants L, Collins G, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG: An International Journal of Obstetrics and Gynaecology*. 2017;124(3):423-432. doi:10.1111/1471-0528.14170

195. Alemayehu D, Chen Y, Markatou M. A comparative study of subgroup identification methods for differential treatment effect: Performance metrics and recommendations. *Statistical Methods in Medical Research*. 2018;27(12):3658-3678. doi:10.1177/0962280217710570

196. Senn S. Statistical pitfalls of personalized medicine. *Nature*. 2018;563(7733):619-621. doi:10.1038/d41586-018-07535-2

197. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005

198. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine*. 2006;144(3):201-209. doi:10.7326/0003-4819-144-3-200602070-00009

199. Gourraud P-A, Henry RG, Cree BAC, Crane JC, Lizee A, Olson MP, Santaniello AV, Datta E, Zhu AH, Bevan CJ, Gelfand JM, Graves JS, Goodin DS, Green AJ, Büdingen HCv, Waubant E, Zamvil SS, Crabtree-Hartman E, Nelson S, Baranzini SE, Hauser SL. Precision medicine in chronic disease management: The multiple sclerosis BioScreen. *Annals of Neurology*. 2014;76(5):633-642. doi:10.1002/ana.24282

200. Stern RH. Individual risk. *Journal of Clinical Hypertension*. 2012;14(4):261-264. doi:10.1111/j.1751-7176.2012.00592.x

201. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, Voysey M, Wharton R, Yu L-M, Moons KG, Altman DG. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*. 2014;14(1):40. doi:10.1186/1471-2288-14-40

202. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, Gagnier J, Borenstein M, Heijden GJMGvd, Dahabreh IJ, Sun X, Sauerbrei W, Walsh M, Ioannidis JPA, Thabane L, Guyatt GH. Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ: Canadian Medical Association Journal*. 2020;192(32):E901-E906. doi:10.1503/cmaj.200077

203. Chakraborty B, Murphy SA. Dynamic treatment regimes. *Annual review of statistics and its application*. 2014;1(1):447-464. doi:10.1146/annurev-statistics-022513-115553

204. Huang Y, Fong Y. Identifying optimal biomarker combinations for treatment selection via a robust kernel method. *Biometrics*. 2014;70(4):891–901. doi:10.1111/biom.12204

Appendix A: R Session Info

```
R version 4.2.0 (2022-04-22 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows Server >= 2012 x64 (build 9200)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252  LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252 LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

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

other attached packages:
 [1] dcurves_0.3.0             TreatmentSelection_2.1.1  cowplot_1.1.1
 [4] ggplotify_0.1.0           gridGraphics_0.5-1        scales_1.2.1
 [7] gridExtra_2.3             rio_0.5.29                riskRegression_2022.09.23
[10] mice_3.14.0               glmnet_4.1-4              Matrix_1.5-1
[13] tictoc_1.1                doParallel_1.0.17         iterators_1.0.14
[16] foreach_1.5.2             trtf_0.4-0                partykit_1.2-16
[19] mvtnorm_1.1-3             libcoin_1.0-9             tram_0.7-2
[22] mlt_1.4-2                 missForest_1.5            survminer_0.4.9
[25] ggpubr_0.4.0              caret_6.0-93              forcats_0.5.2
[28] stringr_1.4.1             dplyr_1.0.10              purrr_0.3.4
[31] readr_2.1.2               tidyr_1.2.1               tibble_3.1.8
[34] tidyverse_1.3.2           grpreg_3.4.0              gtsummary_1.6.2
[37] Hmisc_4.7-1               ggplot2_3.3.6             Formula_1.2-4
[40] lattice_0.20-45           risksetROC_1.0.4.1        MASS_7.3-58.1
[43] survivalROC_1.0.3         basefun_1.1-3             variables_1.1-1
[46] survival_3.4-0            haven_2.5.1

loaded via a namespace (and not attached):
  [1] utf8_1.2.2           rms_6.3-0            tidyselect_1.1.2     htmlwidgets_1.5.4
  [5] pROC_1.18.0          munsell_0.5.0        codetools_0.2-18     interp_1.1-3
  [9] future_1.28.0        withr_2.5.0          colorspace_2.0-3     knitr_1.40
 [13] rstudioapi_0.14      stats4_4.2.0         ggsignif_0.6.3       listenv_0.8.0
 [17] KMsurv_0.1-5         mets_1.3.0           farver_2.1.1         coneproj_1.16
 [21] parallelly_1.32.1    vctrs_0.4.1          generics_0.1.3       TH.data_1.1-1
 [25] ipred_0.9-13         xfun_0.33            itertools_0.1-3      randomForest_4.7-1.1
 [29] R6_2.5.1             timereg_2.0.2        assertthat_0.2.1     multcomp_1.4-20
 [33] nnet_7.3-17          googlesheets4_1.0.1  gtable_0.3.1         globals_0.16.1
 [37] sandwich_3.0-2       timeDate_4021.104    rlang_1.0.6          MatrixModels_0.5-1
 [41] cmprsk_2.2-11        splines_4.2.0        rstatix_0.7.0        ModelMetrics_1.2.2.2
 [45] gargle_1.2.1         broom_1.0.1          checkmate_2.1.0      yaml_2.3.5
 [49] reshape2_1.4.4       abind_1.4-5          modelr_0.1.9         backports_1.4.1
 [53] inum_1.0-4           tools_4.2.0          lava_1.6.10          ellipsis_0.3.2
 [57] RColorBrewer_1.1-3   polynom_1.4-1        Rcpp_1.0.9           plyr_1.8.7
 [61] base64enc_0.1-3      rpart_4.1.16         deldir_1.0-6         zoo_1.8-11
 [65] cluster_2.1.4        fs_1.5.2             magrittr_2.0.3       data.table_1.14.2
 [69] openxlsx_4.2.5       SparseM_1.81         reprex_2.0.2         googledrive_2.0.0
 [73] hms_1.1.2            xtable_1.8-4         jpeg_0.1-9           readxl_1.4.1
 [77] shape_1.4.6          compiler_4.2.0       gt_0.7.0             crayon_1.5.1
 [81] htmltools_0.5.3      tzdb_0.3.0           lubridate_1.8.0      DBI_1.1.3
 [85] dbplyr_2.2.1         broom.helpers_1.9.0  car_3.1-0            cli_3.4.1
 [89] quadprog_1.5-8       gower_1.0.0          pkgconfig_2.0.3      km.ci_0.5-6
 [93] numDeriv_2016.8-1.1  foreign_0.8-82       BB_2019.10-1         binom_1.1-1.1
 [97] recipes_1.0.1        alabama_2022.4-1     xml2_1.3.3           hardhat_1.2.0
[101] rngtools_1.5.2       prodlim_2019.11.13   rvest_1.0.3          yulab.utils_0.0.5
[105] doRNG_1.8.2          digest_0.6.29        cellranger_1.1.0     survMisc_0.5.6
[109] htmlTable_2.4.1      curl_4.3.2           quantreg_5.94        lifecycle_1.0.2
[113] nlme_3.1-159         jsonlite_1.8.0       carData_3.0-5        orthopolynom_1.0-6
[117] fansi_1.0.3          pillar_1.8.1         fastmap_1.1.0        httr_1.4.4
[121] glue_1.6.2           zip_2.2.1            png_0.1-7            class_7.3-20
[125] stringi_1.7.8        polspline_1.1.20     latticeExtra_0.6-30  future.apply_1.9.1
```

Appendix B: Additional Tables

Variable	Linear coefficient
Relapse	
DrugFTY720	-0.83624597
EDSS score (total)	0.08356831
Total volume of Gd-enhanced T1 lesions	0.00027808
Number of relapses in the last 2 years	0.08498321
Number of prior MS treatments	0.00839243
New/enlarging lesions	
DrugFTY720	-0.82102185
Age21-25	0.09824895
Age46-50	-0.0075754
Number of Gd-enhanced T1 lesions	0.00865408
Total volume of Gd-enhanced T1 lesions	0.00010987
Total volume of T2 lesions	7.9628E-06
Duration of MS since 1st symptom	-0.00611288
Quality of Life:Visual analog scale	0.00034345
Lab:Bilirubin (direct/conjugated) BIOCHEM umol/L	0.01109491
Immunosuppressant safety	
DrugFTY720	-0.08582944
SexFemale	0.01653066
Body Mass Index (kg/m^2)	-0.00098386
Age41-45	-0.02337666
EDSS:Brainstem functions	0.0251405
EDSS:Cerebellar functions	-0.0282713
Total volume of Gd-enhanced T1 lesions	1.7187E-05
Prior Glatiramer acetate use=Yes	0.01035474
Prior Natalizumab or other MS treatment use=Yes	-0.03025864
Visual acuity decimal score left	0.0845448
Visual acuity decimal score right	0.21987833
Quality of Life:Mobility	-0.06567299
Quality of Life:Anxiety / Depression	0.01138766
Comedication:Dermatologicals=Yes	0.1159763
Comedication:Genito urinary system and sex hormones=Yes	0.13824079
Comedication:Systemic hormonal preparations, excluding sex hormones and insulins=Yes	0.0814255
Comedication:Various=Yes	0.1643258
Concomitant Disease:Congenital, familial and genetic disorders=Yes	0.09221938
Concomitant Disease:Endocrine disorders=Yes	0.0875366
Concomitant Disease:Gastrointestinal disorders=Yes	0.14388149
Concomitant Disease:Immune system disorders=Yes	0.09874599
Concomitant Disease:Infections and infestations=Yes	0.28316262
Concomitant Disease:Metabolism and nutrition disorders=Yes	0.10159298
Concomitant Disease:Nervous system disorders=Yes	0.00271637
Concomitant Disease:Renal and urinary disorders=Yes	0.02960597
Lab:Alkaline phosphatase, serum BIOCHEM U/L	0.00109141
Lab:Creatinine BIOCHEM umol/L	-0.00219541
Lab:Absolute Lymphocytes HEMA 10E9/L	-0.04312727
Lab:Absolute Neutrophils HEMA 10E9/L	0.0024971
Lab:Mean Cell Volume HEMA fL	-0.00423537
DrugFTY720*Age21-25	-0.10833771
DrugFTY720*Age36-40	0.1322695
DrugFTY720*Age41-45	-0.01964259
DrugFTY720*EDSS:Pyramidal functions	-0.04478977
DrugFTY720*EDSS:Cerebellar functions	-0.00142819
DrugFTY720*Total volume of Gd-enhanced T1 lesions	2.0288E-05
DrugFTY720*Comedication:Musculo-skeletal system=Yes	-0.0433899
DrugFTY720*Comedication:Respiratory system=Yes	0.17076092
DrugFTY720*Comedication:Various=Yes	0.02030373
DrugFTY720*Concomitant Disease:Congenital, familial and genetic disorders=Yes	0.0052955
DrugFTY720*Concomitant Disease:Neoplasms benign, malignant and unspecified (incl cysts and polyps)=Yes	0.02726856
DrugFTY720*Concomitant Disease:Nervous system disorders=Yes	0.13832364
DrugFTY720*Concomitant Disease:Respiratory, thoracic and mediastinal disorders=Yes	0.32565868
DrugFTY720*Lab:Bilirubin (direct/conjugated) BIOCHEM umol/L	-0.00177145
DrugFTY720*Lab:Gamma Glutamyltransferase (GGT) BIOCHEM U/L	-0.00038623
Safety and efficacy	
DrugFTY720	-0.5591239
EDSS:Cerebellar functions	0.01634399
DrugFTY720*EDSS:Cerebellar functions	-0.00579678
EDSS:Bowel and bladder functions	0.02290899
DrugFTY720*EDSS:Bowel and bladder functions	-0.00117191
EDSS:Cerebral (or mental) functions	0.00142342
DrugFTY720*EDSS:Cerebral (or mental) functions	0.00058645
EDSS score (total)	0.00318225
DrugFTY720*EDSS score (total)	-0.00067104
Number of Gd-enhanced T1 lesions	0.00335139
DrugFTY720*Number of Gd-enhanced T1 lesions	-0.00320212

Variable	Linear coefficient
Total volume of Gd-enhanced T1 lesions	0.00048238
DrugFTY720*Total volume of Gd-enhanced T1 lesions	-0.0003717
Number of relapses in the last 2 years	0.00495667
DrugFTY720*Number of relapses in the last 2 years	-0.00072299
Number of prior MS treatments	0.05699765
DrugFTY720*Number of prior MS treatments	0.02557324
Quality of Life:Mobility	0.01072284
DrugFTY720*Quality of Life:Mobility	-4.1307E-05
Comedication:Nervous system=Yes	0.06065012
DrugFTY720*Comedication:Nervous system=Yes	0.02342168
Comedication:Various=Yes	0.00297254
DrugFTY720*Comedication:Various=Yes	0.1404988
Concomitant Disease:Musculoskeletal and connective tissue disorders=Yes	0.08000896
DrugFTY720*Concomitant Disease:Musculoskeletal and connective tissue disorders=Yes	-0.00687186

Table A1 *Regression coefficients* Coefficients in the final models by the outcomes for which a regression model was chosen as the best performing method.

Day	Relapse	T2 MRI	Immune safety	Composite
1	0.001039235	0	0.012176397	0.007753065
2	0.005265211	0	0.030720957	0.013342154
3	0.007393432	0	0.036978594	0.015587643
4	0.010584958	0	0.044809043	0.018967465
5	0.011656296	0	0.062300264	0.021227794
6	0.013802592	0	0.073547114	0.023494985
7	0.016940398	0	0.078418162	0.028046449
8	0.021264894	0	0.081675686	0.03265238
9	0.022354955	0	0.086557171	0.033810936
10	0.023446849	0	0.096420195	0.034971777
11	0.024542384	0	0.106279142	0.036137383
12	0.024542384	0	0.111218315	0.036137383
13	0.025640657	0	0.117879614	0.037306444
14	0.025640657	0	0.124597379	0.037306444
15	0.025640657	0	0.137937011	0.037306444
16	0.026739701	0	0.148235708	0.039648871
17	0.027825912	0	0.155152822	0.040821759
18	0.028926622	0	0.160368307	0.041995702
19	0.032241464	0	0.167339549	0.046706766
20	0.032241464	0	0.174382362	0.046706766
21	0.033337259	0.001504197	0.184871324	0.047889131
22	0.035549945	0.001504197	0.190230946	0.051457777
23	0.037789644	0.001504197	0.197392765	0.053851256
24	0.038912351	0.001504197	0.206472335	0.055050591
25	0.038912351	0.001504197	0.211953408	0.055050591
26	0.040036955	0.001504197	0.220918392	0.058657636
27	0.043418884	0.001504197	0.233796698	0.062279241
28	0.043418884	0.001504197	0.239407995	0.062279241
29	0.043418884	0.001504197	0.245046691	0.062279241
30	0.045685734	0.001504197	0.246931781	0.065913172
31	0.045685734	0.003016845	0.246931781	0.065913172
32	0.047966061	0.003016845	0.250698411	0.068349078
33	0.049111057	0.003016845	0.254459011	0.069570744
34	0.049111057	0.003016845	0.264038197	0.070794359
35	0.049111057	0.003016845	0.265940736	0.070794359
36	0.05255885	0.003016845	0.269802277	0.075715332
37	0.054867984	0.003016845	0.273676394	0.078190493
38	0.054867984	0.003016845	0.273676394	0.078190493
39	0.057187283	0.003016845	0.275593502	0.081922629
40	0.058351559	0.003016845	0.279439158	0.083176406
41	0.058351559	0.003016845	0.281392718	0.083176406

Day	Relapse	T2 MRI	Immune safety	Composite
42	0.058351559	0.003016845	0.283349272	0.083176406
43	0.058351559	0.003016845	0.293165471	0.08443234
44	0.05951751	0.003016845	0.293165471	0.086949348
45	0.05951751	0.003016845	0.299115263	0.086949348
46	0.05951751	0.003016845	0.303097323	0.086949348
47	0.05951751	0.003016845	0.311098291	0.086949348
48	0.060684865	0.003016845	0.315117166	0.08947391
49	0.06183978	0.003016845	0.319148358	0.090739079
50	0.063010683	0.003016845	0.319148358	0.092005992
51	0.065342121	0.003016845	0.323193759	0.094544478
52	0.065342121	0.003016845	0.325222841	0.094544478
53	0.066518026	0.004516483	0.335408787	0.097091836
54	0.066518026	0.004516483	0.34572146	0.097091836
55	0.067695851	0.004516483	0.34572146	0.098367694
56	0.068875406	0.004516483	0.34572146	0.09964621
57	0.068875406	0.004516483	0.347794564	0.09964621
58	0.068875406	0.004516483	0.354032844	0.09964621
59	0.070058174	0.004516483	0.364360331	0.100928988
60	0.070058174	0.004516483	0.368577201	0.100928988
61	0.072414351	0.004516483	0.370690624	0.103504473
62	0.073412968	0.004516483	0.381278146	0.10479648
63	0.074606855	0.004516483	0.387666831	0.107394875
64	0.075777295	0.004516483	0.387666831	0.108701558
65	0.075777295	0.004516483	0.394104059	0.110011438
66	0.078185136	0.004516483	0.394104059	0.112635633
67	0.079377175	0.004516483	0.398430461	0.113951113
68	0.079377175	0.004516483	0.402770471	0.113951113
69	0.081797521	0.004516483	0.404945963	0.116587664
70	0.081797521	0.004516483	0.40712577	0.116587664
71	0.081797521	0.004516483	0.413659867	0.116587664
72	0.081797521	0.004516483	0.415692772	0.116587664
73	0.0854491	0.004516483	0.420098324	0.120557363
74	0.087890229	0.004516483	0.4223073	0.123211778
75	0.087890229	0.00604244	0.426710781	0.123211778
76	0.09033665	0.00604244	0.428932284	0.125872737
77	0.091562724	0.00604244	0.435620595	0.127206892
78	0.092776012	0.00755234	0.44007284	0.128543643
79	0.092776012	0.00755234	0.442319409	0.128543643
80	0.094007303	0.00755234	0.444569684	0.129882958
81	0.094007303	0.00755234	0.449080912	0.131224711
82	0.096476984	0.00755234	0.449080912	0.133915552

Day	Relapse	T2 MRI	Immune safety	Composite
83	0.097715218	0.00755234	0.451341986	0.135264426
84	0.097715218	0.00755234	0.462674851	0.137968956
85	0.097715218	0.00755234	0.467245255	0.137968956
86	0.097715218	0.00755234	0.469539413	0.140690282
87	0.097715218	0.00755234	0.481074745	0.140690282
88	0.098956531	0.00755234	0.488052136	0.142055671
89	0.101430317	0.00755234	0.497419325	0.147540439
90	0.10267841	0.00755234	0.502100977	0.148916102
91	0.105181015	0.00755234	0.50914958	0.153059123
92	0.105181015	0.00755234	0.511527977	0.160003444
93	0.105181015	0.00755234	0.518670399	0.168408773
94	0.105181015	0.009086933	0.518670399	0.169815975
95	0.105181015	0.009086933	0.521070214	0.169815975
96	0.105181015	0.010606351	0.523475203	0.169815975
97	0.105181015	0.010606351	0.523475203	0.169815975
98	0.105181015	0.012146688	0.5306861	0.172635433
99	0.105181015	0.012146688	0.533108633	0.176886671
100	0.105181015	0.012146688	0.535543769	0.178308163
101	0.106442907	0.012146688	0.535543769	0.181158599
102	0.106442907	0.012146688	0.537985039	0.181158599
103	0.106442907	0.012146688	0.542882493	0.182587801
104	0.107707067	0.012146688	0.545338498	0.185451441
105	0.108974534	0.012146688	0.547774217	0.185451441
106	0.110242924	0.012146688	0.552700062	0.18688542
107	0.111512473	0.013696308	0.555155737	0.188320895
108	0.111512473	0.013696308	0.557647179	0.188320895
109	0.111512473	0.013696308	0.557647179	0.189758394
110	0.114055242	0.015249722	0.557647179	0.192638755
111	0.115328899	0.015249722	0.560142537	0.194081594
112	0.115328899	0.015249722	0.562641829	0.196972593
113	0.115328899	0.015249722	0.565145407	0.196972593
114	0.116605234	0.015249722	0.567654448	0.198420891
115	0.119152049	0.015249722	0.570170421	0.201332817
116	0.119152049	0.015249722	0.572691836	0.201332817
117	0.119152049	0.015249722	0.572691836	0.201332817
118	0.120439195	0.015249722	0.575219179	0.202794255
119	0.121728885	0.015249722	0.577752776	0.204257656
120	0.121728885	0.015249722	0.580292262	0.204257656
121	0.121728885	0.015249722	0.580292262	0.204257656
122	0.121728885	0.015249722	0.582836425	0.205722627
123	0.121728885	0.015249722	0.58538492	0.205722627

Day	Relapse	T2 MRI	Immune safety	Composite
124	0.121728885	0.015249722	0.58538492	0.205722627
125	0.121728885	0.015249722	0.587938309	0.205722627
126	0.121728885	0.016806454	0.590496434	0.205722627
127	0.121728885	0.016806454	0.595627622	0.207189097
128	0.121728885	0.016806454	0.595627622	0.208657299
129	0.123020095	0.016806454	0.598200815	0.210127802
130	0.123020095	0.016806454	0.608516693	0.210127802
131	0.123020095	0.016806454	0.611117992	0.210127802
132	0.123020095	0.016806454	0.611117992	0.210127802
133	0.123020095	0.016806454	0.613727143	0.210127802
134	0.123020095	0.016806454	0.613727143	0.210127802
135	0.124313081	0.016806454	0.613727143	0.211600349
136	0.125591424	0.016806454	0.616349789	0.213074813
137	0.130771661	0.016806454	0.616349789	0.218991924
138	0.130771661	0.016806454	0.62161075	0.218991924
139	0.131913382	0.018369385	0.62161075	0.218991924
140	0.131913382	0.018369385	0.624255419	0.218991924
141	0.133219576	0.018369385	0.626906199	0.221966324
142	0.134349719	0.018369385	0.629529594	0.224952676
143	0.134349719	0.018369385	0.632191563	0.224952676
144	0.135644997	0.018369385	0.634859343	0.226451048
145	0.135644997	0.018369385	0.634859343	0.226451048
146	0.135644997	0.018369385	0.642857297	0.226451048
147	0.138258419	0.018369385	0.648235375	0.229457065
148	0.138258419	0.019936037	0.653648096	0.229457065
149	0.14089917	0.019936037	0.653648096	0.232475688
150	0.142223644	0.019936037	0.661814924	0.233989296
151	0.142223644	0.019936037	0.661814924	0.233989296
152	0.142223644	0.019936037	0.664551159	0.233989296
153	0.146195976	0.019936037	0.675529158	0.238550399
154	0.146195976	0.019936037	0.677920144	0.238550399
155	0.148850306	0.021485927	0.686265809	0.241604708
156	0.148850306	0.024631999	0.689058563	0.241604708
157	0.150189527	0.026208685	0.689058563	0.24313722
158	0.151532115	0.02778843	0.697467103	0.244673371
159	0.151532115	0.02778843	0.703103596	0.244673371
160	0.152876998	0.02778843	0.70589697	0.246211865
161	0.152876998	0.030955825	0.70589697	0.246211865
162	0.154224236	0.030955825	0.711587107	0.247753167
163	0.155507045	0.034132478	0.711587107	0.250849286
164	0.155507045	0.034132478	0.717270393	0.250849286

Day	Relapse	T2 MRI	Immune safety	Composite
165	0.156859781	0.038891066	0.723013789	0.252403106
166	0.156859781	0.046892126	0.728745591	0.252403106
167	0.156859781	0.051702365	0.734442752	0.253960258
168	0.158215606	0.063010588	0.737366954	0.255519816
169	0.163664151	0.087682633	0.749149415	0.261783765
170	0.166390234	0.10770596	0.752112654	0.263357615
171	0.167749739	0.10770596	0.761045787	0.264934518
172	0.167749739	0.109408648	0.761045787	0.264934518
173	0.167749739	0.11622925	0.761045787	0.264934518
174	0.167749739	0.138714973	0.767036474	0.268095569
175	0.167749739	0.156216223	0.770041023	0.268095569
176	0.167749739	0.199427282	0.773028332	0.269682981
177	0.169131934	0.232914745	0.778947759	0.272863832
178	0.170515902	0.257958765	0.787989802	0.27605544
179	0.174681897	0.28952233	0.794128182	0.280864382
180	0.176074727	0.311837357	0.809514529	0.282471717
181	0.176074727	0.336575567	0.815748964	0.282471717
182	0.176074727	0.372727868	0.81887762	0.285693802
183	0.18025732	0.437479046	0.825113655	0.293789793
184	0.184479031	0.489273382	0.831436298	0.300318569
185	0.184479031	0.506271167	0.837791198	0.300318569
186	0.184479031	0.538596582	0.837791198	0.303599861
187	0.185892525	0.561475299	0.837791198	0.305244351
188	0.185892525	0.587615135	0.84097877	0.305244351
189	0.187309322	0.614423717	0.84417449	0.308544014
190	0.187309322	0.670020369	0.850589151	0.311858682
191	0.187309322	0.7018128	0.850589151	0.311858682
192	0.187309322	0.728565791	0.853767666	0.311858682
193	0.188729287	0.746650796	0.853767666	0.313521152
194	0.188729287	0.765104629	0.853767666	0.313521152
195	0.188729287	0.777486701	0.860244106	0.313521152
196	0.188729287	0.783794836	0.860244106	0.313521152
197	0.188729287	0.799145349	0.863496364	0.315191
198	0.188729287	0.808790066	0.863496364	0.315191
199	0.188729287	0.815224644	0.863496364	0.315191
200	0.188729287	0.828298925	0.863496364	0.315191
201	0.188729287	0.834891297	0.863496364	0.315191
202	0.19015496	0.8448357	0.863496364	0.31853893
203	0.191582994	0.84816748	0.863496364	0.31853893
204	0.191582994	0.84816748	0.863496364	0.31853893
205	0.191582994	0.84816748	0.863496364	0.31853893

Day	Relapse	T2 MRI	Immune safety	Composite
206	0.191582994	0.851513044	0.863496364	0.31853893
207	0.191582994	0.854871521	0.863496364	0.31853893
208	0.193014421	0.854871521	0.863496364	0.320217559
209	0.194447909	0.854871521	0.870030399	0.321898728
210	0.195883177	0.854871521	0.870030399	0.321898728
211	0.197321548	0.854871521	0.87331145	0.32358263
212	0.197321548	0.854871521	0.879904695	0.32358263
213	0.198763304	0.85820022	0.883174788	0.325270051
214	0.198763304	0.85820022	0.883174788	0.325270051
215	0.20154676	0.85820022	0.883174788	0.32865382
216	0.20154676	0.85820022	0.883174788	0.32865382
217	0.20154676	0.85820022	0.886492771	0.32865382
218	0.202995645	0.85820022	0.88981788	0.33035234
219	0.202995645	0.85820022	0.893162357	0.33035234
220	0.202995645	0.864984098	0.896517372	0.33035234
221	0.204448341	0.86835763	0.896517372	0.332053639
222	0.205905036	0.86835763	0.896517372	0.335469126
223	0.205905036	0.86835763	0.899839322	0.335469126
224	0.207365826	0.86835763	0.90316921	0.337183896
225	0.208828874	0.86835763	0.90316921	0.337183896
226	0.210294297	0.875252575	0.906548454	0.338902423
227	0.211763034	0.875252575	0.91329034	0.340624539
228	0.213234717	0.875252575	0.920068619	0.340624539
229	0.214710024	0.875252575	0.920068619	0.344081195
230	0.214710024	0.875252575	0.93037878	0.345815709
231	0.214710024	0.875252575	0.933836816	0.345815709
232	0.214710024	0.875252575	0.940743768	0.345815709
233	0.216188927	0.875252575	0.940743768	0.347553719
234	0.216188927	0.875252575	0.944236693	0.349295304
235	0.216188927	0.875252575	0.947697227	0.349295304
236	0.216188927	0.875252575	0.951212064	0.351039737
237	0.217671612	0.875252575	0.951212064	0.352788171
238	0.219139433	0.875252575	0.957699866	0.354541095
239	0.219139433	0.875252575	0.957699866	0.354541095
240	0.219139433	0.875252575	0.961242093	0.354541095
241	0.222100438	0.875252575	0.964749839	0.358058776
242	0.222100438	0.875252575	0.964749839	0.358058776
243	0.222100438	0.875252575	0.964749839	0.359827677
244	0.225074616	0.875252575	0.968316106	0.363375066
245	0.225074616	0.875252575	0.968316106	0.363375066
246	0.225074616	0.875252575	0.979072509	0.363375066

Day	Relapse	T2 MRI	Immune safety	Composite
247	0.225074616	0.875252575	0.979072509	0.363375066
248	0.226579141	0.875252575	0.982639996	0.365159857
249	0.228086586	0.875252575	0.982639996	0.366947964
250	0.228086586	0.875252575	0.982639996	0.368738998
251	0.228086586	0.875252575	0.98626444	0.368738998
252	0.231108453	0.875252575	0.993507317	0.370533862
253	0.232623143	0.875252575	0.997167901	0.370533862
254	0.232623143	0.875252575	0.997167901	0.370533862
255	0.234139731	0.875252575	0.997167901	0.375940638
256	0.235659144	0.875252575	0.997167901	0.379562523
257	0.240236633	0.875252575	0.997167901	0.385025554
258	0.241769104	0.875252575	0.997167901	0.388689548
259	0.244828279	0.875252575	1.000838515	0.392374391
260	0.244828279	0.878730327	1.008208019	0.394227114
261	0.244828279	0.878730327	1.008208019	0.394227114
262	0.244828279	0.878730327	1.011907276	0.394227114
263	0.244828279	0.878730327	1.011907276	0.396082407
264	0.244828279	0.878730327	1.015617679	0.396082407
265	0.247922282	0.878730327	1.015617679	0.401665414
266	0.247922282	0.878730327	1.03043503	0.401665414
267	0.249472427	0.878730327	1.045438395	0.407280711
268	0.251024904	0.878730327	1.053081265	0.409160448
269	0.251024904	0.878730327	1.056919331	0.409160448
270	0.252580877	0.878730327	1.060768805	0.411044696
271	0.252580877	0.878730327	1.07238314	0.411044696
272	0.25414037	0.88223318	1.076227044	0.414828002
273	0.25414037	0.88223318	1.076227044	0.416726806
274	0.25414037	0.885744415	1.076227044	0.416726806
275	0.255704218	0.885744415	1.080129686	0.422457228
276	0.257272901	0.885744415	1.080129686	0.42437396
277	0.257272901	0.885744415	1.080129686	0.42437396
278	0.260415987	0.885744415	1.080129686	0.428215062
279	0.263530732	0.885744415	1.080129686	0.432070657
280	0.263530732	0.885744415	1.080129686	0.432070657
281	0.263530732	0.885744415	1.084050663	0.434005956
282	0.265112996	0.885744415	1.084050663	0.435945969
283	0.266698904	0.885744415	1.087983798	0.437890296
284	0.266698904	0.885744415	1.091931795	0.437890296
285	0.26828748	0.885744415	1.091931795	0.439838162
286	0.26828748	0.885744415	1.095900281	0.439838162
287	0.26828748	0.885744415	1.095900281	0.439838162

Day	Relapse	T2 MRI	Immune safety	Composite
288	0.26828748	0.885744415	1.095900281	0.439838162
289	0.26828748	0.885744415	1.095900281	0.439838162
290	0.269858205	0.885744415	1.095900281	0.441793846
291	0.271455629	0.885744415	1.095900281	0.443752993
292	0.271455629	0.885744415	1.095900281	0.443752993
293	0.273056728	0.885744415	1.095900281	0.445716653
294	0.273056728	0.885744415	1.095900281	0.447683952
295	0.274641814	0.885744415	1.099878955	0.451630569
296	0.274641814	0.885744415	1.099878955	0.451630569
297	0.274641814	0.885744415	1.103868968	0.451630569
298	0.274641814	0.885744415	1.103868968	0.451630569
299	0.274641814	0.885744415	1.107871891	0.453611137
300	0.274641814	0.885744415	1.107871891	0.453611137
301	0.274641814	0.885744415	1.107871891	0.453611137
302	0.276250848	0.885744415	1.11188725	0.455596701
303	0.276250848	0.885744415	1.11188725	0.455596701
304	0.277864	0.885744415	1.11188725	0.455596701
305	0.282723545	0.885744415	1.11188725	0.461578939
306	0.282723545	0.885744415	1.11188725	0.461578939
307	0.285977242	0.885744415	1.11188725	0.465588901
308	0.285977242	0.885744415	1.11188725	0.467599905
309	0.287588406	0.885744415	1.11188725	0.469615626
310	0.287588406	0.885744415	1.11188725	0.469615626
311	0.289224368	0.885744415	1.11188725	0.471636696
312	0.289224368	0.885744415	1.115874532	0.471636696
313	0.290864489	0.885744415	1.119921539	0.473662536
314	0.290864489	0.885744415	1.123983652	0.473662536
315	0.290864489	0.885744415	1.128012268	0.473662536
316	0.290864489	0.885744415	1.136211398	0.475692517
317	0.290864489	0.885744415	1.140332704	0.475692517
318	0.290864489	0.885744415	1.140332704	0.475692517
319	0.292508914	0.885744415	1.144469314	0.477726998
320	0.294157427	0.885744415	1.144469314	0.479767252
321	0.294157427	0.885744415	1.144469314	0.479767252
322	0.294157427	0.885744415	1.144469314	0.479767252
323	0.295809753	0.885744415	1.1486199	0.481812894
324	0.297466309	0.889313572	1.1486199	0.485919243
325	0.297466309	0.889313572	1.1486199	0.485919243
326	0.297466309	0.889313572	1.152796442	0.485919243
327	0.297466309	0.889313572	1.161136706	0.485919243
328	0.297466309	0.889313572	1.161136706	0.485919243

Day	Relapse	T2 MRI	Immune safety	Composite
329	0.299126267	0.889313572	1.16535217	0.487979032
330	0.299126267	0.889313572	1.16535217	0.487979032
331	0.300791977	0.889313572	1.16535217	0.490047014
332	0.300791977	0.889313572	1.16535217	0.490047014
333	0.300791977	0.889313572	1.178079537	0.492118032
334	0.300791977	0.889313572	1.178079537	0.492118032
335	0.302440257	0.889313572	1.182350148	0.494193819
336	0.302440257	0.889313572	1.186633294	0.494193819
337	0.302440257	0.889313572	1.190934131	0.494193819
338	0.302440257	0.889313572	1.190934131	0.494193819
339	0.302440257	0.889313572	1.190934131	0.494193819
340	0.302440257	0.889313572	1.190934131	0.494193819
341	0.302440257	0.889313572	1.195250675	0.494193819
342	0.305801916	0.889313572	1.208177003	0.49836661
343	0.305801916	0.889313572	1.216947363	0.500460833
344	0.305801916	0.889313572	1.216947363	0.500460833
345	0.305801916	0.892908536	1.221249897	0.500460833
346	0.307489706	0.903765512	1.221249897	0.50255819
347	0.307489706	0.903765512	1.225676078	0.50255819
348	0.307489706	0.907362304	1.234523434	0.50255819
349	0.310877755	0.911015491	1.234523434	0.50677065
350	0.310877755	0.925500009	1.243539023	0.508891537
351	0.314264049	0.93290925	1.243539023	0.513143242
352	0.317684341	0.940193624	1.243539023	0.517415221
353	0.317684341	0.943950185	1.243539023	0.517415221
354	0.317684341	0.962808709	1.248082538	0.517415221
355	0.317684341	0.966609433	1.257225728	0.517415221
356	0.319380374	0.970420807	1.261820482	0.517415221
357	0.321101697	0.985829645	1.271048545	0.523859231
358	0.322830001	1.021170992	1.285074625	0.532543416
359	0.322830001	1.045326162	1.285074625	0.532543416
360	0.322830001	1.049413074	1.289791115	0.534724492
361	0.322830001	1.057654191	1.289791115	0.536910148
362	0.322830001	1.057654191	1.289791115	0.539107224
363	0.322830001	1.065970734	1.289791115	0.539107224
364	0.322830001	1.091295851	1.299309969	0.539107224
365	0.322830001	1.12133766	1.299309969	0.539107224
366	0.322830001	1.143389338	1.299309969	0.539107224
367	0.322830001	1.156951215	1.299309969	0.541320694
368	0.322830001	1.161510142	1.299309969	0.543540742
369	0.322830001	1.170682582	1.299309969	0.543540742

Day	Relapse	T2 MRI	Immune safety	Composite
370	0.322830001	1.17987703	1.299309969	0.543540742
371	0.324581072	1.189127069	1.308938522	0.545773181
372	0.324581072	1.193821011	1.308938522	0.545773181
373	0.326340041	1.203251213	1.313797902	0.548011208
374	0.326340041	1.203251213	1.318675047	0.548011208
375	0.326340041	1.203251213	1.318675047	0.548011208
376	0.328101914	1.207989074	1.323568061	0.550254702
377	0.329846251	1.207989074	1.328475898	0.552505193
378	0.329846251	1.207989074	1.328475898	0.552505193
379	0.329846251	1.212743739	1.328475898	0.552505193
380	0.329846251	1.212743739	1.333402958	0.552505193
381	0.331616085	1.212743739	1.333402958	0.554761289
382	0.331616085	1.212743739	1.333402958	0.554761289
383	0.333388533	1.217538274	1.333402958	0.55702118
384	0.333388533	1.217538274	1.333402958	0.55702118
385	0.333388533	1.217538274	1.338363639	0.559284778
386	0.333388533	1.217538274	1.338363639	0.559284778
387	0.333388533	1.217538274	1.338363639	0.559284778
388	0.333388533	1.217538274	1.338363639	0.559284778
389	0.333388533	1.217538274	1.338363639	0.559284778
390	0.333388533	1.222363676	1.338363639	0.559284778
391	0.333388533	1.222363676	1.343347672	0.559284778
392	0.335165201	1.222363676	1.343347672	0.56155545
393	0.336946238	1.222363676	1.348351862	0.563833329
394	0.336946238	1.222363676	1.348351862	0.563833329
395	0.336946238	1.222363676	1.348351862	0.563833329
396	0.336946238	1.222363676	1.353374732	0.563833329
397	0.336946238	1.222363676	1.358355549	0.563833329
398	0.336946238	1.222363676	1.358355549	0.563833329
399	0.336946238	1.222363676	1.358355549	0.563833329
400	0.336946238	1.222363676	1.358355549	0.563833329
401	0.336946238	1.222363676	1.358355549	0.563833329
402	0.336946238	1.222363676	1.363419852	0.563833329
403	0.336946238	1.222363676	1.368506322	0.563833329
404	0.336946238	1.222363676	1.368506322	0.563833329
405	0.336946238	1.222363676	1.378615462	0.563833329
406	0.336946238	1.222363676	1.378615462	0.563833329
407	0.338731179	1.222363676	1.378615462	0.56611975
408	0.338731179	1.222363676	1.378615462	0.56611975
409	0.338731179	1.222363676	1.378615462	0.56611975
410	0.338731179	1.222363676	1.378615462	0.56611975

Day	Relapse	T2 MRI	Immune safety	Composite
411	0.338731179	1.227240993	1.378615462	0.56611975
412	0.338731179	1.232143297	1.383783007	0.56611975
413	0.338731179	1.237071694	1.394180523	0.56611975
414	0.340522152	1.237071694	1.394180523	0.568416112
415	0.340522152	1.237071694	1.399413337	0.568416112
416	0.344097344	1.24202705	1.399413337	0.573029843
417	0.344097344	1.24202705	1.404600585	0.573029843
418	0.344097344	1.24202705	1.404600585	0.573029843
419	0.344097344	1.24202705	1.404600585	0.573029843
420	0.344097344	1.24202705	1.404600585	0.575355927
421	0.344097344	1.24202705	1.404600585	0.575355927
422	0.344097344	1.24202705	1.404600585	0.575355927
423	0.344097344	1.24202705	1.404600585	0.575355927
424	0.344097344	1.24202705	1.409806012	0.575355927
425	0.344097344	1.24202705	1.409806012	0.575355927
426	0.344097344	1.24202705	1.42033441	0.575355927
427	0.344097344	1.24202705	1.425657267	0.575355927
428	0.345908603	1.24202705	1.425657267	0.577687933
429	0.345908603	1.24202705	1.425657267	0.577687933
430	0.347723514	1.24202705	1.431002547	0.58002564
431	0.349540962	1.24202705	1.436369089	0.582367654
432	0.349540962	1.24202705	1.436369089	0.582367654
433	0.349540962	1.24202705	1.44717616	0.582367654
434	0.349540962	1.24202705	1.44717616	0.582367654
435	0.349540962	1.24202705	1.44717616	0.582367654
436	0.349540962	1.24202705	1.452580359	0.582367654
437	0.349540962	1.24202705	1.452580359	0.582367654
438	0.351366891	1.24202705	1.452580359	0.584723651
439	0.351366891	1.24202705	1.452580359	0.584723651
440	0.353198244	1.24202705	1.452580359	0.587087042
441	0.353198244	1.24202705	1.452580359	0.587087042
442	0.353198244	1.246965142	1.452580359	0.587087042
443	0.353198244	1.246965142	1.458074538	0.589456183
444	0.353198244	1.246965142	1.458074538	0.589456183
445	0.353198244	1.246965142	1.458074538	0.589456183
446	0.353198244	1.246965142	1.463616446	0.589456183
447	0.353198244	1.246965142	1.469208236	0.589456183
448	0.353198244	1.246965142	1.469208236	0.589456183
449	0.353198244	1.246965142	1.469208236	0.594217305
450	0.353198244	1.246965142	1.486155319	0.596604051
451	0.353198244	1.246965142	1.497600524	0.596604051

Day	Relapse	T2 MRI	Immune safety	Composite
452	0.355038208	1.246965142	1.503359232	0.601393464
453	0.355038208	1.246965142	1.503359232	0.601393464
454	0.355038208	1.246965142	1.51494939	0.601393464
455	0.355038208	1.246965142	1.51494939	0.601393464
456	0.355038208	1.246965142	1.51494939	0.603795761
457	0.355038208	1.246965142	1.520787767	0.60620327
458	0.355038208	1.246965142	1.526655834	0.608618916
459	0.355038208	1.246965142	1.532545913	0.608618916
460	0.355038208	1.246965142	1.532545913	0.608618916
461	0.358708259	1.246965142	1.532545913	0.611042884
462	0.358708259	1.246965142	1.544333922	0.613473304
463	0.358708259	1.246965142	1.544333922	0.613473304
464	0.358708259	1.246965142	1.544333922	0.613473304
465	0.358708259	1.246965142	1.544333922	0.613473304
466	0.358708259	1.246965142	1.544333922	0.615911185
467	0.358708259	1.246965142	1.544333922	0.615911185
468	0.358708259	1.246965142	1.544333922	0.615911185
469	0.358708259	1.246965142	1.544333922	0.615911185
470	0.358708259	1.246965142	1.544333922	0.615911185
471	0.36056573	1.246965142	1.544333922	0.615911185
472	0.36056573	1.246965142	1.544333922	0.615911185
473	0.36056573	1.246965142	1.544333922	0.615911185
474	0.36056573	1.246965142	1.544333922	0.615911185
475	0.36056573	1.246965142	1.544333922	0.615911185
476	0.36056573	1.246965142	1.544333922	0.615911185
477	0.36056573	1.246965142	1.550313423	0.615911185
478	0.36056573	1.246965142	1.550313423	0.615911185
479	0.36056573	1.246965142	1.550313423	0.615911185
480	0.362426366	1.246965142	1.550313423	0.618356306
481	0.364292895	1.246965142	1.550313423	0.62327117
482	0.364292895	1.246965142	1.550313423	0.62327117
483	0.364292895	1.252038374	1.550313423	0.62327117
484	0.364292895	1.252038374	1.550313423	0.62327117
485	0.364292895	1.252038374	1.550313423	0.62327117
486	0.366165027	1.252038374	1.550313423	0.625737275
487	0.366165027	1.252038374	1.550313423	0.625737275
488	0.366165027	1.252038374	1.550313423	0.625737275
489	0.366165027	1.252038374	1.556324584	0.625737275
490	0.366165027	1.252038374	1.556324584	0.625737275
491	0.368000487	1.252038374	1.562360965	0.625737275
492	0.368000487	1.252038374	1.562360965	0.625737275

Day	Relapse	T2 MRI	Immune safety	Composite
493	0.368000487	1.252038374	1.574506003	0.633180412
494	0.369888036	1.252038374	1.574506003	0.635675067
495	0.371778023	1.252038374	1.574506003	0.63817466
496	0.371778023	1.252038374	1.574506003	0.63817466
497	0.371778023	1.252038374	1.574506003	0.63817466
498	0.373672365	1.252038374	1.574506003	0.640680849
499	0.373672365	1.252038374	1.580619617	0.640680849
500	0.373672365	1.252038374	1.586767256	0.640680849
501	0.373672365	1.252038374	1.586767256	0.643193551
502	0.373672365	1.252038374	1.598990682	0.643193551
503	0.373672365	1.252038374	1.598990682	0.643193551
504	0.373672365	1.252038374	1.605219484	0.643193551
505	0.373672365	1.252038374	1.605219484	0.643193551
506	0.373672365	1.252038374	1.611398208	0.643193551
507	0.373672365	1.252038374	1.623988046	0.643193551
508	0.373672365	1.252038374	1.623988046	0.643193551
509	0.373672365	1.252038374	1.623988046	0.643193551
510	0.373672365	1.252038374	1.623988046	0.643193551
511	0.375572996	1.252038374	1.623988046	0.645712207
512	0.375572996	1.252038374	1.623988046	0.645712207
513	0.375572996	1.252038374	1.623988046	0.645712207
514	0.375572996	1.252038374	1.623988046	0.645712207
515	0.377455038	1.252038374	1.623988046	0.648238821
516	0.377455038	1.252038374	1.623988046	0.648238821
517	0.377455038	1.252038374	1.623988046	0.648238821
518	0.379366596	1.252038374	1.623988046	0.650773542
519	0.381284035	1.252038374	1.623988046	0.653315966
520	0.381284035	1.252038374	1.623988046	0.655866199
521	0.383208338	1.252038374	1.623988046	0.658425132
522	0.383208338	1.252038374	1.623988046	0.658425132
523	0.383208338	1.252038374	1.623988046	0.658425132
524	0.383208338	1.252038374	1.623988046	0.658425132
525	0.385115058	1.252038374	1.623988046	0.663566236
526	0.387050741	1.252038374	1.630350127	0.666146478
527	0.387050741	1.252038374	1.630350127	0.666146478
528	0.387050741	1.252038374	1.630350127	0.666146478
529	0.387050741	1.252038374	1.630350127	0.668735279
530	0.387050741	1.252038374	1.649634474	0.668735279
531	0.388990927	1.252038374	1.656134245	0.67133062
532	0.388990927	1.252038374	1.662666162	0.67133062
533	0.390934919	1.252038374	1.669225852	0.673932364

Day	Relapse	T2 MRI	Immune safety	Composite
534	0.392860708	1.252038374	1.669225852	0.67654247
535	0.392860708	1.252038374	1.669225852	0.679161386
536	0.392860708	1.252038374	1.675735004	0.679161386
537	0.394817545	1.252038374	1.675735004	0.681789588
538	0.394817545	1.252038374	1.682359313	0.681789588
539	0.394817545	1.252038374	1.695707841	0.681789588
540	0.394817545	1.252038374	1.695707841	0.681789588
541	0.394817545	1.252038374	1.695707841	0.684430016
542	0.394817545	1.252038374	1.695707841	0.687076943
543	0.394817545	1.252038374	1.695707841	0.687076943
544	0.394817545	1.252038374	1.695707841	0.687076943
545	0.394817545	1.252038374	1.702467903	0.689735606
546	0.396784769	1.252038374	1.702467903	0.692399352
547	0.396784769	1.252038374	1.702467903	0.69507846
548	0.396784769	1.252038374	1.702467903	0.69507846
549	0.396784769	1.252038374	1.716087075	0.69507846
550	0.396784769	1.252038374	1.716087075	0.69507846
551	0.396784769	1.252038374	1.716087075	0.69507846
552	0.396784769	1.252038374	1.716087075	0.69507846
553	0.396784769	1.252038374	1.716087075	0.69507846
554	0.398761632	1.252038374	1.716087075	0.697765017
555	0.400745263	1.252038374	1.716087075	0.700461001
556	0.402734305	1.252038374	1.716087075	0.703165992
557	0.402734305	1.252038374	1.716087075	0.703165992
558	0.402734305	1.252038374	1.716087075	0.703165992
559	0.402734305	1.252038374	1.716087075	0.703165992
560	0.404727676	1.252038374	1.716087075	0.705878093
561	0.404727676	1.252038374	1.716087075	0.708597553
562	0.404727676	1.252038374	1.716087075	0.708597553
563	0.404727676	1.252038374	1.729845149	0.708597553
564	0.404727676	1.252038374	1.729845149	0.708597553
565	0.40672493	1.252038374	1.729845149	0.714059944
566	0.40672493	1.252038374	1.729845149	0.714059944
567	0.40672493	1.252038374	1.729845149	0.714059944
568	0.40672493	1.252038374	1.729845149	0.714059944
569	0.40672493	1.252038374	1.729845149	0.714059944
570	0.40672493	1.252038374	1.729845149	0.714059944
571	0.40672493	1.252038374	1.729845149	0.716802641
572	0.40872669	1.252038374	1.73677909	0.719550539
573	0.40872669	1.252038374	1.73677909	0.722303762
574	0.40872669	1.257222162	1.73677909	0.722303762

Day	Relapse	T2 MRI	Immune safety	Composite
575	0.40872669	1.257222162	1.73677909	0.722303762
576	0.40872669	1.257222162	1.73677909	0.725063353
577	0.40872669	1.257222162	1.743740625	0.725063353
578	0.40872669	1.257222162	1.743740625	0.725063353
579	0.40872669	1.257222162	1.750728824	0.725063353
580	0.40872669	1.257222162	1.757747989	0.725063353
581	0.40872669	1.257222162	1.757747989	0.725063353
582	0.410741553	1.257222162	1.771875839	0.727829643
583	0.410741553	1.257222162	1.771875839	0.727829643
584	0.410741553	1.257222162	1.771875839	0.727829643
585	0.410741553	1.257222162	1.771875839	0.727829643
586	0.410741553	1.257222162	1.778983827	0.727829643
587	0.414783544	1.257222162	1.778983827	0.73338739
588	0.414783544	1.257222162	1.778983827	0.73338739
589	0.414783544	1.257222162	1.778983827	0.736178065
590	0.414783544	1.257222162	1.786132246	0.736178065
591	0.41681368	1.257222162	1.793324155	0.738976601
592	0.41681368	1.257222162	1.793324155	0.738976601
593	0.41681368	1.257222162	1.793324155	0.738976601
594	0.41681368	1.257222162	1.800560785	0.738976601
595	0.41681368	1.257222162	1.800560785	0.738976601
596	0.41681368	1.262418607	1.800560785	0.738976601
597	0.41681368	1.262418607	1.800560785	0.738976601
598	0.41681368	1.262418607	1.800560785	0.738976601
599	0.41681368	1.262418607	1.807838473	0.738976601
600	0.418851762	1.262418607	1.807838473	0.74178634
601	0.418851762	1.262418607	1.807838473	0.74178634
602	0.418851762	1.262418607	1.815153302	0.74178634
603	0.418851762	1.262418607	1.815153302	0.74178634
604	0.418851762	1.262418607	1.829922545	0.74178634
605	0.418851762	1.262418607	1.837373513	0.74178634
606	0.418851762	1.262418607	1.837373513	0.74178634
607	0.418851762	1.262418607	1.844856643	0.74178634
608	0.418851762	1.262418607	1.844856643	0.74178634
609	0.418851762	1.262418607	1.852375459	0.74178634
610	0.418851762	1.262418607	1.852375459	0.74178634
611	0.420897222	1.262418607	1.852375459	0.74178634
612	0.42295629	1.262418607	1.859842027	0.744605332
613	0.425021056	1.262418607	1.859842027	0.744605332
614	0.425021056	1.262418607	1.867448844	0.744605332
615	0.425021056	1.262418607	1.875107341	0.744605332

Day	Relapse	T2 MRI	Immune safety	Composite
616	0.425021056	1.262418607	1.875107341	0.744605332
617	0.425021056	1.262418607	1.875107341	0.747432104
618	0.429119545	1.262418607	1.882869933	0.75312615
619	0.429119545	1.262418607	1.882869933	0.75312615
620	0.429119545	1.262418607	1.890750864	0.75312615
621	0.43121284	1.262418607	1.890750864	0.75312615
622	0.43121284	1.262418607	1.890750864	0.75312615
623	0.43121284	1.262418607	1.890750864	0.75312615
624	0.43121284	1.262418607	1.890750864	0.75312615
625	0.43121284	1.262418607	1.890750864	0.75312615
626	0.43121284	1.262418607	1.890750864	0.75312615
627	0.43121284	1.262418607	1.890750864	0.75312615
628	0.435378046	1.262418607	1.890750864	0.758877161
629	0.435378046	1.262418607	1.890750864	0.758877161
630	0.439590939	1.262418607	1.890750864	0.767572526
631	0.4417188	1.262418607	1.890750864	0.770487779
632	0.4417188	1.262418607	1.890750864	0.770487779
633	0.4417188	1.262418607	1.890750864	0.770487779
634	0.4417188	1.262418607	1.890750864	0.773410795
635	0.443852947	1.262418607	1.890750864	0.77634388
636	0.44596667	1.262418607	1.890750864	0.782241869
637	0.44596667	1.262418607	1.890750864	0.782241869
638	0.44596667	1.262418607	1.890750864	0.782241869
639	0.448116796	1.262418607	1.898676541	0.782241869
640	0.448116796	1.262418607	1.898676541	0.782241869
641	0.448116796	1.262418607	1.906647397	0.785224098
642	0.450284625	1.262418607	1.906647397	0.788216548
643	0.450284625	1.262418607	1.906647397	0.788216548
644	0.450284625	1.262418607	1.906647397	0.791219481
645	0.450284625	1.262418607	1.906647397	0.791219481
646	0.450284625	1.262418607	1.906647397	0.791219481
647	0.450284625	1.262418607	1.906647397	0.791219481
648	0.450284625	1.262418607	1.906647397	0.791219481
649	0.452477027	1.262418607	1.906647397	0.794245094
650	0.452477027	1.262418607	1.906647397	0.794245094
651	0.452477027	1.262418607	1.906647397	0.794245094
652	0.454674152	1.267708091	1.906647397	0.797280753
653	0.454674152	1.267708091	1.906647397	0.797280753
654	0.454674152	1.267708091	1.906647397	0.797280753
655	0.456856232	1.267708091	1.906647397	0.797280753
656	0.456856232	1.267708091	1.906647397	0.797280753

Day	Relapse	T2 MRI	Immune safety	Composite
657	0.456856232	1.267708091	1.906647397	0.797280753
658	0.456856232	1.267708091	1.906647397	0.797280753
659	0.456856232	1.267708091	1.906647397	0.797280753
660	0.456856232	1.267708091	1.906647397	0.797280753
661	0.456856232	1.267708091	1.906647397	0.797280753
662	0.456856232	1.267708091	1.906647397	0.797280753
663	0.458963501	1.267708091	1.906647397	0.800328819
664	0.461186363	1.267708091	1.914660187	0.803385735
665	0.461186363	1.267708091	1.914660187	0.803385735
666	0.461186363	1.267708091	1.914660187	0.803385735
667	0.463414857	1.267708091	1.914660187	0.803385735
668	0.463414857	1.267708091	1.914660187	0.803385735
669	0.465650377	1.267708091	1.914660187	0.8064491
670	0.467889365	1.267708091	1.914660187	0.809520641
671	0.470133917	1.267708091	1.914660187	0.809520641
672	0.470133917	1.267708091	1.914660187	0.809520641
673	0.470133917	1.267708091	1.914660187	0.809520641
674	0.470133917	1.267708091	1.914660187	0.812600604
675	0.470133917	1.267708091	1.914660187	0.812600604
676	0.470133917	1.267708091	1.914660187	0.812600604
677	0.470133917	1.267708091	1.914660187	0.812600604
678	0.470133917	1.267708091	1.914660187	0.812600604
679	0.472389678	1.267708091	1.914660187	0.815688536
680	0.472389678	1.267708091	1.914660187	0.815688536
681	0.472389678	1.267708091	1.914660187	0.818784604
682	0.472389678	1.267708091	1.914660187	0.818784604
683	0.474649699	1.267708091	1.930831193	0.821887916
684	0.474649699	1.267708091	1.938989681	0.825000403
685	0.474649699	1.267708091	1.938989681	0.825000403
686	0.474649699	1.267708091	1.947202673	0.825000403
687	0.476914745	1.267708091	1.947202673	0.828124411
688	0.476914745	1.267708091	1.947202673	0.828124411
689	0.481475805	1.267708091	1.955502282	0.834425215
690	0.483766865	1.267708091	1.963848263	0.83759157
691	0.486061325	1.267708091	1.963848263	0.8407646
692	0.486061325	1.267708091	1.963848263	0.8407646
693	0.486061325	1.267708091	1.97223691	0.8407646
694	0.486061325	1.273046032	1.97223691	0.8407646
695	0.488371479	1.273046032	1.97223691	0.843960571
696	0.490691628	1.273046032	1.97223691	0.843960571
697	0.490691628	1.278397546	1.97223691	0.843960571

Day	Relapse	T2 MRI	Immune safety	Composite
698	0.490691628	1.278397546	1.97223691	0.843960571
699	0.490691628	1.278397546	1.97223691	0.843960571
700	0.490691628	1.278397546	1.97223691	0.843960571
701	0.492968314	1.278397546	1.97223691	0.843960571
702	0.495314903	1.278397546	1.97223691	0.843960571
703	0.495314903	1.278397546	1.97223691	0.843960571
704	0.495314903	1.278397546	1.97223691	0.843960571
705	0.495314903	1.2837635	1.97223691	0.843960571
706	0.495314903	1.2837635	1.97223691	0.847278613
707	0.495314903	1.294867812	1.97223691	0.847278613
708	0.495314903	1.30612756	1.989699157	0.847278613
709	0.495314903	1.30612756	1.989699157	0.847278613
710	0.495314903	1.31775835	1.989699157	0.847278613
711	0.497818714	1.31775835	1.989699157	0.847278613
712	0.500328999	1.323712747	1.998566942	0.850729612
713	0.500328999	1.329729929	1.998566942	0.850729612
714	0.500328999	1.335739127	1.998566942	0.850729612
715	0.500328999	1.354250597	1.998566942	0.850729612
716	0.500328999	1.366852275	2.007902547	0.850729612
717	0.500328999	1.373238679	2.017421827	0.850729612
718	0.500328999	1.373238679	2.037076935	0.854385644
719	0.500328999	1.373238679	2.037076935	0.854385644
720	0.500328999	1.379642124	2.037076935	0.854385644

Table A2 *Baseline hazard* Baseline cumulative hazard functions from the final models by the outcomes for which a regression model was chosen as the best performing method.

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety	Composite
Month 6						
Transformation tree	0.491	0.495	0.527	0.504	0.552	0.484
Transformation forest	0.670	0.652	0.699	0.498	0.597	0.620
Elastic net	0.669	0.653	0.537	0.449	0.582	0.587
Grouped lasso	0.672	0.656	0.526	0.434	0.575	0.604
Month 12						
Transformation tree	0.508	0.475	0.535	0.527	0.529	0.498
Transformation forest	0.614	0.706	0.658	0.563	0.579	0.554
Elastic net	0.684	0.730	0.565	0.530	0.597	0.612
Grouped lasso	0.685	0.721	0.554	0.528	0.584	0.623
Month 24						
Transformation tree	0.491	0.447	0.550	0.504	0.529	0.490
Transformation forest	0.647	0.689	0.660	0.565	0.617	0.604
Elastic net	0.714	0.753	0.581	0.547	0.634	0.641
Grouped lasso	0.721	0.745	0.573	0.544	0.644	0.652

Table A3 *Cross-validated monthly area under the curve* Cumulative time-dependent area under the curve at 6, 12, and 24 months estimated via cross-validation in the development dataset. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

Method	Relapse	T2 MRI	3m CDP	Safety	Immune safety	Composite
Month 6						
Transformation tree	0.164	0.197	0.066	0.049	0.248	0.205
Transformation forest	0.143	0.165	0.065	0.049	0.242	0.188
Elastic net	0.138	0.166	0.066	0.049	0.244	0.187
Grouped lasso	0.137	0.169	0.068	0.049	0.263	0.185
Month 12						
Transformation tree	0.233	0.357	0.123	0.081	0.236	0.26
Transformation forest	0.214	0.234	0.12	0.08	0.232	0.254
Elastic net	0.192	0.222	0.123	0.08	0.231	0.236
Grouped lasso	0.194	0.237	0.127	0.08	0.257	0.235
Month 24						
Transformation tree	0.276	0.322	0.164	0.127	0.177	0.264
Transformation forest	0.24	0.227	0.159	0.125	0.172	0.25
Elastic net	0.213	0.205	0.164	0.124	0.171	0.233
Grouped lasso	0.213	0.223	0.172	0.125	0.2	0.233

Table A4 *Cross-validated monthly Brier score* Average time dependent Brier score at 6, 12, and 24 months estimated via cross-validation in the development dataset. T2 MRI: New/enlarging lesions, 3m CDP: Confirmed disability progression, Immune safety: Immunosuppressant safety, Composite: Safety and efficacy.

Transformation tree	Transformation forest	Elastic net	Grouped lasso
Relapse			
Total volume of Gd-enhanced T1 lesions	Concomitant Disease: Metabolism and nutrition disorders	_DrugFTY720.Age36-40	Age21-25
	Quality of Life: Mobility	_DrugFTY720.Comedication: Cardiovascular system=Yes	Age26-30
	Total volume of Gd-enhanced T1 lesions	_DrugFTY720.Concomitant Disease: Metabolism and nutrition disorders=Yes	Age31-35
	Total volume of T2 lesions	_DrugFTY720.Number of prior MS treatments	Age36-40
		Comedication: Blood and blood forming organs=Yes	Age41-45
		Concomitant Disease: Metabolism and nutrition disorders=Yes	Age46-50
		EDSS score (total)	Age51-55
		EDSS: Bowel and bladder functions	Comedication: Blood and blood forming organs=Yes
		EDSS: Pyramidal functions	Comedication: Cardiovascular system=Yes
		Number of prior MS treatments	Comedication: Dermatologicals=Yes
		Number of relapses in the last 2 years	Comedication: Musculo-skeletal system=Yes
		Total volume of Gd-enhanced T1 lesions	Comedication: Nervous system=Yes
		Total volume of T2 lesions	Comedication: Respiratory system=Yes
			Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes
			Concomitant Disease: Endocrine disorders=Yes
			Concomitant Disease: Gastrointestinal disorders=Yes
			Concomitant Disease: General disorders and administration site conditions=Yes
			Concomitant Disease: Infections and infestations=Yes
			Concomitant Disease: Metabolism and nutrition disorders=Yes
			Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes
			Concomitant Disease: Nervous system disorders=Yes
			Concomitant Disease: Psychiatric disorders=Yes
			Concomitant Disease: Renal and urinary disorders=Yes
			Concomitant Disease: Skin and subcutaneous tissue disorders=Yes
			Duration of MS since 1st symptom
			EDSS score (total)
			EDSS: Bowel and bladder functions

Transformation tree	Transformation forest	Elastic net	Grouped lasso
			EDSS: Cerebellar functions
			EDSS: Pyramidal functions
			Lab: Absolute Neutrophils HEMA 10E9/L
			Lab: SGOT (AST) BIOCHEM U/L
			Number of Gd-enhanced T1 lesions
			Number of months since recent relapse
			Number of prior MS treatments
			Number of relapses in the last 2 years
			Prior Glatiramer acetate use=Yes
			Prior Interferon beta use=Yes
			Prior Natalizumab or other MS treatment use=Yes
			Total volume of Gd-enhanced T1 lesions
			Total volume of T2 lesions
New/enlarging lesions			
Number of Gd-enhanced T1 lesions	Age	_DrugFTY720.Concomitant Disease: Gastrointestinal disorders=Yes	Comedication: Alimentary tract and metabolism=Yes
Total volume of Gd-enhanced T1 lesions	Number of Gd-enhanced T1 lesions	Age21-25	Concomitant Disease: Endocrine disorders=Yes
	Total volume of Gd-enhanced T1 lesions	Age46-50	Concomitant Disease: Infections and infestations=Yes
	Total volume of T1 hypointense lesions	Age51-55	Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes
	Total volume of T2 lesions	Duration of MS since 1st symptom	Duration of MS since 1st symptom
		Lab: Bilirubin (direct/conjugated) BIOCHEM umol/L	Lab: Bilirubin (direct/conjugated) BIOCHEM umol/L
		Number of Gd-enhanced T1 lesions	Number of Gd-enhanced T1 lesions
		Number of months since recent relapse	Quality of Life: Visual analog scale
		Quality of Life: Usual activities	Total volume of Gd-enhanced T1 lesions
		Quality of Life: Visual analog scale	Total volume of T2 lesions
		Total volume of Gd-enhanced T1 lesions	
		Total volume of T2 lesions	
Confirmed disability progression			
	Concomitant Disease: Musculoskeletal and connective tissue disorders	_DrugFTY720.Comedication: Cardiovascular system=Yes	Comedication: Cardiovascular system=Yes
	EDSS score (total)	_DrugFTY720.Comedication: Various=Yes	Comedication: Nervous system=Yes
	EDSS: Cerebral (or mental) functions	_DrugFTY720.Concomitant Disease: Metabolism and nutrition disorders=Yes	Comedication: Respiratory system=Yes

Transformation tree	Transformation forest	Elastic net	Grouped lasso
	MSFC: Mean of 9-hole peg test	_DrugFTY720.Concomitant Disease: Skin and subcutaneous tissue disorders=Yes	Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes
		_DrugFTY720.EDSS score (total)	Comedication: Various=Yes
		_DrugFTY720.Number of prior MS treatments	Concomitant Disease: Congenital, familial and genetic disorders=Yes
		_DrugFTY720.Prior Interferon beta use=Yes	Concomitant Disease: Endocrine disorders=Yes
		Age41-45	Concomitant Disease: Gastrointestinal disorders=Yes
		Comedication: Nervous system=Yes	Concomitant Disease: Investigations=Yes
		Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes	Concomitant Disease: Metabolism and nutrition disorders=Yes
		Comedication: Various=Yes	Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes
		Concomitant Disease: Endocrine disorders=Yes	Concomitant Disease: Nervous system disorders=Yes
		Concomitant Disease: Gastrointestinal disorders=Yes	Concomitant Disease: Skin and subcutaneous tissue disorders=Yes
		Concomitant Disease: Investigations=Yes	EDSS score (total)
		Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes	EDSS: Cerebral (or mental) functions
		Concomitant Disease: Nervous system disorders=Yes	Lab: Absolute Neutrophils HEMA 10E9/L
		Concomitant Disease: Respiratory, thoracic and mediastinal disorders=Yes	Lab: Albumin BIOCHEM g/L
		Lab: Absolute Neutrophils HEMA 10E9/L	MSFC: Mean of 9-hole peg test
		Lab: Albumin BIOCHEM g/L	Number of prior MS treatments
		Lab: White Blood Cell (total) HEMA 10E9/L	Quality of Life: Anxiety / Depression
		MSFC: Mean of 9-hole peg test	Quality of Life: Mobility
		Number of prior MS treatments	
		Quality of Life: Anxiety / Depression	
		Quality of Life: Mobility	
		Quality of Life: Visual analog scale	
		Total volume of T1 hypointense lesions	
		Visual acuity decimal score left	
		Visual acuity decimal score right	
Safety			
Concomitant Disease: Gastrointestinal disorders	Concomitant Disease: Gastrointestinal disorders	_DrugFTY720.EDSS: Cerebral (or mental) functions	Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes
	Concomitant Disease: Metabolism and nutrition disorders	Age51-55	Concomitant Disease: Gastrointestinal disorders=Yes

Transformation tree	Transformation forest	Elastic net	Grouped lasso
	Lab: Creatinine BIOCHEM umol/L	Concomitant Disease: Gastrointestinal disorders=Yes	Concomitant Disease: Metabolism and nutrition disorders=Yes
	Quality of Life: Self-care	Quality of Life: Anxiety / Depression	Concomitant Disease: Nervous system disorders=Yes
		Total volume of T2 lesions	EDSS: Bowel and bladder functions
			EDSS: Cerebral (or mental) functions
			Lab: Absolute Eosinophils HEMA 10E9/L
			Lab: Absolute Monocytes HEMA 10E9/L
			Total volume of T2 lesions
Immunosuppressant safety			
Comedication: Genito urinary system and sex hormones		_DrugFTY720.Age21-25	Age21-25
		_DrugFTY720.Age36-40	Age26-30
		_DrugFTY720.Comedication: Musculo-skeletal system=Yes	Age31-35
		_DrugFTY720.Comedication: Respiratory system=Yes	Age36-40
		_DrugFTY720.Comedication: Various=Yes	Age41-45
		_DrugFTY720.Concomitant Disease: Congenital, familial and genetic disorders=Yes	Age46-50
		_DrugFTY720.Concomitant Disease: Immune system disorders=Yes	Age51-55
		_DrugFTY720.Concomitant Disease: Nervous system disorders=Yes	Comedication: Blood and blood forming organs=Yes
		_DrugFTY720.Concomitant Disease: Respiratory, thoracic and mediastinal disorders=Yes	Comedication: Cardiovascular system=Yes
		_DrugFTY720.EDSS: Cerebellar functions	Comedication: Dermatologicals=Yes
		_DrugFTY720.EDSS: Pyramidal functions	Comedication: Genito urinary system and sex hormones=Yes
		_DrugFTY720.Total volume of Gd-enhanced T1 lesions	Comedication: Respiratory system=Yes
		Age36-40	Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes
		Age41-45	Comedication: Various=Yes
		Comedication: Dermatologicals=Yes	Concomitant Disease: Congenital, familial and genetic disorders=Yes
		Comedication: Genito urinary system and sex hormones=Yes	Concomitant Disease: Endocrine disorders=Yes
		Comedication: Systemic hormonal preparations, excluding sex hormones and insulins=Yes	Concomitant Disease: Eye disorders=Yes
		Comedication: Various=Yes	Concomitant Disease: Gastrointestinal disorders=Yes
		Concomitant Disease: Congenital, familial and genetic disorders=Yes	Concomitant Disease: Immune system disorders=Yes

Transformation tree	Transformation forest	Elastic net	Grouped lasso
		Concomitant Disease: Endocrine disorders=Yes	Concomitant Disease: Infections and infestations=Yes
		Concomitant Disease: Gastrointestinal disorders=Yes	Concomitant Disease: Metabolism and nutrition disorders=Yes
		Concomitant Disease: Immune system disorders=Yes	Concomitant Disease: Neoplasms benign, malignant and unspecified (incl cysts and polyps)=Yes
		Concomitant Disease: Infections and infestations=Yes	Concomitant Disease: Nervous system disorders=Yes
		Concomitant Disease: Metabolism and nutrition disorders=Yes	Concomitant Disease: Psychiatric disorders=Yes
		Concomitant Disease: Nervous system disorders=Yes	Concomitant Disease: Renal and urinary disorders=Yes
		Concomitant Disease: Renal and urinary disorders=Yes	Concomitant Disease: Reproductive system and breast disorders=Yes
		EDSS: Brainstem functions	Concomitant Disease: Respiratory, thoracic and mediastinal disorders=Yes
		EDSS: Cerebellar functions	EDSS: Bowel and bladder functions
		Lab: Absolute Lymphocytes HEMA 10E9/L	EDSS: Brainstem functions
		Lab: Absolute Neutrophils HEMA 10E9/L	EDSS: Cerebellar functions
		Lab: Alkaline phosphatase, serum BIOCHEM U/L	EDSS: Pyramidal functions
		Lab: Creatinine BIOCHEM umol/L	Lab: Absolute Lymphocytes HEMA 10E9/L
		Lab: Mean Cell Volume HEMA fL	Lab: Absolute Neutrophils HEMA 10E9/L
		Prior Glatiramer acetate use=Yes	Lab: Alkaline phosphatase, serum BIOCHEM U/L
		Prior Natalizumab or other MS treatment use=Yes	Lab: Creatinine BIOCHEM umol/L
		Quality of Life: Mobility	Lab: Gamma Glutamyltransferase (GGT) BIOCHEM U/L
		SexFemale	Lab: Mean Cell Volume HEMA fL
		Total volume of Gd-enhanced T1 lesions	Number of Gd-enhanced T1 lesions
		Visual acuity decimal score left	Prior Glatiramer acetate use=Yes
		Visual acuity decimal score right	Prior Natalizumab or other MS treatment use=Yes
			Quality of Life: Mobility
			Racenon.Caucasian
			SexFemale
			Total volume of Gd-enhanced T1 lesions
			Visual acuity decimal score left
			Visual acuity decimal score right
Safety and efficacy			

Transformation tree	Transformation forest	Elastic net	Grouped lasso
Total volume of Gd-enhanced T1 lesions	Concomitant Disease: Metabolism and nutrition disorders	_DrugFTY720.Age36-40	Comedication: Blood and blood forming organs=Yes
	Total volume of Gd-enhanced T1 lesions	_DrugFTY720.Comedication: Various=Yes	Comedication: Nervous system=Yes
		Comedication: Blood and blood forming organs=Yes	Comedication: Respiratory system=Yes
		Comedication: Nervous system=Yes	Comedication: Various=Yes
		Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes	Concomitant Disease: Musculoskeletal and connective tissue disorders=Yes
		EDSS: Bowel and bladder functions	EDSS score (total)
		EDSS: Cerebellar functions	EDSS: Bowel and bladder functions
		EDSS: Cerebral (or mental) functions	EDSS: Cerebellar functions
		Number of prior MS treatments	EDSS: Cerebral (or mental) functions
		Number of relapses in the last 2 years	Number of Gd-enhanced T1 lesions
		Quality of Life: Mobility	Number of prior MS treatments
		Total volume of Gd-enhanced T1 lesions	Number of relapses in the last 2 years
		Total volume of T2 lesions	Quality of Life: Mobility
			Total volume of Gd-enhanced T1 lesions
			Total volume of T2 lesions

Table A5 *Important variables* List of variables deemed important at least in two folds of the five-fold cross-validation by being selected in the model or, in the case of forest, by having a permutation importance greater than that expected from random variation. All lists omit the drug because it was included in the models by design. Also, the lists for the grouped lasso method omits the interaction terms because the method simultaneously selects the main terms and their interaction with treatment by design. Green the predictor is deemed important by all four methods; Yellow the predictor is deemed important by three methods; Gray the predictor is deemed important by two methods.

Outcome	Proportion recommended treatment	Average benefit of no treatment in those recommended none	Average benefit of treatment in those recommended	Decrease in rate of outcomes under marker-based treatment	Variance in estimated treatment effect	Total gain
Relapse	1 (1-1)	0 (0-0)	0.223 (0.154-0.296)	0 (0-0)	0.001 (0-0.01)	0.029 (0.016-0.096)
New/enlarging lesions	1 (1-1)	0 (0-0)	0.259 (0.188-0.333)	0 (0-0)	0.001 (0-0.011)	0.034 (0.003-0.103)
Confirmed disability progression	0.516 (0.478-0.555)	0.007 (-0.146-0.172)	-0.032 (-0.153-0.09)	0.004 (-0.07-0.085)	0.006 (0.004-0.014)	0.061 (0.054-0.099)
Safety	0.488 (0.452-0.527)	0 (-0.084-0.094)	-0.083 (-0.172-0.005)	0 (-0.042-0.049)	0.009 (0.007-0.017)	0.075 (0.064-0.109)
Immunosuppressant safety	0.489 (0.454-0.526)	0.077 (0.006-0.147)	-0.045 (-0.139-0.045)	0.039 (0.003-0.075)	0.005 (0.002-0.015)	0.063 (0.036-0.112)
Safety and efficacy	1 (1-1)	0 (0-0)	0.089 (0.011-0.169)	0 (0-0)	0.012 (0.002-0.034)	0.103 (0.033-0.18)

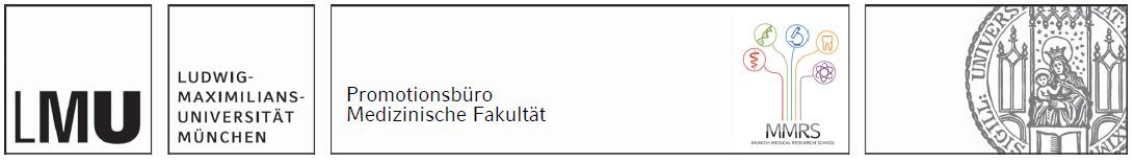
Table A6 *Measures of predicted treatment response* Treatment effect measures derived from actual and counterfactual predictions from the final model in the external validation dataset, with their uncertainty (95% confidence interval).

Acknowledgements

First and foremost, I am thankful to my supervisors. Prof. Dr. Ulrich Mansmann was an invaluable guide and had faith in me every step of the way. Prof. Dr. Martin Kerschensteiner and Dr. Heidi Seibold were always understanding of the challenges in the project and provided useful insights. I am also thankful to the PhD program coordinators, Dr. Annette Hartmann and Monika Darchinger, who patiently responded to my questions and supported my progress. I would also like to acknowledge formal and informal exchanges with all the other PhD students and colleagues who indirectly influenced me during this PhD. Last but not least, I am grateful to my family, partner, and my friends for their never-ending support.

This thesis was only possible because Novartis, the sponsor of the FREEDOMS trials, provided the anonymous study data via the Clinical Study Data Request data sharing platform. I am thankful to all Novartis employees involved in the process. I would also like to acknowledge the contributions of Prof. Dr. Ulrike Held, my dear friend Dr. Kelly Reeve, and Dr. Joachim Havla in conceptualization of the data access application. Also detrimental to the success of this project was the diligent work of Mr. Josef Herker, who prepared the analysis dataset.

Affidavit



Affidavit

Ön, Begüm Irmak
_____
Surname, first name

Marchioninistr. 15
_____
Street

81377, München, Germany
_____
Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Prediction of prognosis and response to fingolimod in people with relapsing-remitting multiple sclerosis

…….

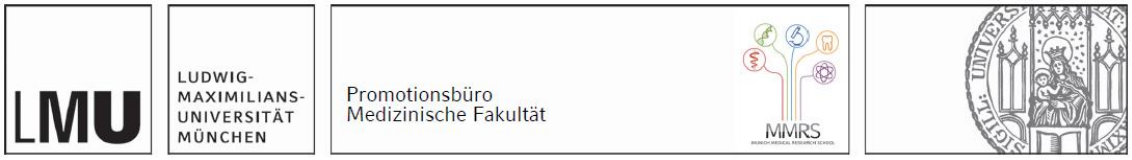
is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

München, 18.12.2023 Begüm Irmak Ön

_____ _____
place, date Signature doctoral candidate

Confirmation of congruency



Confirmation of congruency between printed and electronic version of the doctoral thesis

Ön, Begüm Irmak
Surname, first name

Marchioninistr. 15
Street

81377, München, Germany
Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Prediction of prognosis and response to fingolimod in people with relapsing-remitting multiple sclerosis

……

is congruent with the printed version both in content and format.

München, 18.12.2023 — Begüm Irmak Ön

place, date — Signature doctoral candidate

List of publications

Published manuscripts

Reeve K, **On BI**, Havla J, Burns J, Gosteli-Peter MA, Alabsawi A, Alayash Z, Götschi A, Seibold H, Mansmann U, Held U. Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. *Cochrane Database of Systematic Reviews*. 2023;2023(9)doi:10.1002/14651858.CD013606.pub2

Seker BIO, Reeve K, Havla J, Burns J, Gosteli MA, Lutterotti A, Schippling S, Mansmann U, Held U. (Protocol) Prognostic models for predicting clinical disease progression, worsening and activity in people with multiple sclerosis. *Cochrane Database of Systematic Reviews*. 2020;(5) doi:10.1002/14651858.CD013606

Hapfelmeier A, **On BI**, Mühlau M, Kirschke JS, Berthele A, Gasperi C, Mansmann U, Wuschek A, Bussas M, Boeker M, Bayas A, Senel M, Havla J, Kowarik MC, Kuhn K, Gatz I, Spengler H, Wiestler B, Grundl L, Sepp D, Hemmer B. Retrospective cohort study to devise a treatment decision score predicting adverse 24-month radiological activity in early multiple sclerosis. *Therapeutic Advances in Neurological Disorders*. 2023;16:1-25. doi:10.1177/17562864231161892

Submitted manuscripts

Sakr AM, Mansmann U, Havla J, **Ön BI**. “Framework for Personalized Prediction of Treatment Response in Relapsing-Remitting Multiple Sclerosis: A Replication Study in Independent Data”

Under review by *BMC Medical Research Methodology*

Manuscripts in preparation

Buchka S, **Ön BI**, Havla J, Mansmann U. “Individual surrogacy of MRI T2 lesion information for future disease severity within recent MS Phase II and III trials: A multi-trial synthesis.”