MSJ Experimental Translational Clinical

Original Research Article

Overall and patient-level comparative effectiveness of dimethyl fumarate and fingolimod: A precision medicine application to the Observatoire Français de la Sclérose en Plaques registry

Multiple Sclerosis Journal—Experimental, Translational and Clinical

July-September 2022, 1–11

DOI: 10.1177/20552173221116591

Gabrielle Simoneau, Xiaotong Jiang, Fabien Rollot, Lu Tian, Massimiliano Copetti, Matthieu Guéry, Marta Ruiz, Sandra Vukusic, Carl de Moor, Fabio Pellegrini and on behalf of OFSEP investigators

Abstract

Background: Comparing real-world effectiveness and tolerability of therapies for relapsing-remitting multiple sclerosis is increasingly important, though average treatment effects fail to capture possible treatment effect heterogeneity. With the clinical course of the disease being highly heterogeneous across patients, precision medicine methods enable treatment response heterogeneity investigations.

Objective: To compare real-world effectiveness and discontinuation profiles between dimethyl fumarate and fingolimod while investigating treatment effect heterogeneity with precision medicine methods.

Methods: Adults initiating dimethyl fumarate or fingolimod as a second-line therapy were selected from a French registry. The primary outcome was annualized relapse rate at 12 months. Seven secondary outcomes relative to discontinuation and disease progression were considered. A precision medicine framework was used to characterize treatment effect heterogeneity.

Results: Annualized relapse rates at 12 months were similar for dimethyl fumarate and fingolimod. The odd of treatment persistence was 47% lower for patients treated with dimethyl fumarate relative to those treated with fingolimod (odds ratio: 0.53, 95% confidence interval: 0.39, 0.70). None of the five precision medicine scoring approaches identified treatment heterogeneity.

Conclusion: These findings substantiated the similar effectiveness and different discontinuation profiles for dimethyl fumarate and fingolimod as a second-line therapy for relapsing-remitting multiple sclerosis, with no significant effect heterogeneity observed.

Keywords: Comparative effectiveness, dimethyl fumarate, fingolimod, precision medicine, propensity score, registry

Date received: 17 November 2021; accepted 12 July 2022

Introduction

Multiple sclerosis (MS) is a chronic neurological disease characterized by an episodic course of relapse and recovery, eventually leading to progressive disease, long-term disability, and loss of quality of life for most patients. Though etiology and prevention remain poorly understood, several disease-modifying therapies (DMTs) have been approved to treat relapsing-remitting MS (RRMS), the most frequent initial presentation of the disease.[1] Nonetheless, the clinical course of RRMS patients is highly variable, mitigated by prognostic factors and heterogeneous response to treatments.[2] Observational studies, which are increasingly important for comparing the effectiveness of DMTs in real-world settings, typically focus on estimating average treatment effects. However, this approach fails to account for heterogeneous treatment responses and does not address the clinical imperative

Correspondence to:
Gabrielle Simoneau,
Biogen Canada, 3250 Bloor St W, Mississauga, Toronto, Canada, M8X 2X9.
gabrielle.simoneau@biogen.com

Gabrielle Simoneau,
Biogen Canada, Mississauga, ON, Canada

Xiaotong Jiang,
Biogen Inc., Cambridge, MA, USA

Fabien Rollot,
Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France
Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, Hôpital Pierre Wertheimer, Hospices Civils de Lyon, Bron, France
Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, Lyon, France
Eugéne Devic EDMUS Foundation Against Multiple Sclerosis, State-Approved Foundation, Bron, France

Lu Tian,
Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA

Massimiliano Copetti,
Unit of Biostatistics, Fondazione IRCCS Casa Sollievo della Sofferenza Hospital, San Giovanni Rotondo, Italy

Matthieu Guéry,
Marta Ruiz,
Biogen France SAS, Paris, France

Sandra Vukusic,
Université Claude Bernard Lyon 1, Université de Lyon, Lyon, France
Service de Neurologie, sclérose en plaques, pathologies de la myéline et neuro-inflammation, Hôpital Pierre Wertheimer, Hospices Civils de Lyon, Bron, France
Observatoire Français de la Sclérose en Plaques, Centre de Recherche en Neurosciences de Lyon, INSERM 1028 et CNRS UMR 5292, Lyon, France
Eugéne Devic EDMUS Foundation Against Multiple Sclerosis, State-Approved Foundation, Bron, France

Carl de Moor,
Biogen Inc, Cambridge, MA, USA

Fabio Pellegrini,
Global Medical, Biogen Idec Inc, Cambridge, MA, USA

to find "the right drug for the right patient at the right time."

Novel approaches to approximate an individualized treatment response (ITR) score from observational data can help identify subgroups more likely to benefit from one treatment over another and facilitate precision medicine.[3,4] Compared to traditional post hoc subgroup analyses,[5] the ITR score methodology has three main advantages. First, the ITR score is a continuous score that accounts for multiple baseline covariates while traditional subgroup analyses are typically applied separately by covariate. Second, methods based on continuous scores have been shown to be more powerful than traditional subgroup analyses in detecting treatment effect heterogeneity.[6] Third, ITR scores directly investigate the interaction between treatment and covariates and allow to explore different subgrouping. In the RRMS literature, ITR methods have been applied to examine responses to dimethyl fumarate (DMF) in clinical trials and observational contexts.[3,4] In both contexts, there were signals suggestive of differential responses to DMF, underscoring the need to examine treatment effect heterogeneity when comparing the effectiveness and tolerability of DMTs.

DMF and fingolimod (FTY) are two DMTs approved for RRMS. A 2015 network meta-analysis of clinical trials found that FTY was among the top performing therapies for preventing the recurrence of relapses during the first 24 months of treatment.[1] However, strong assertions about the comparative efficacy of some therapies, including DMF, were limited by a lack of quality evidence. Subgroup analyses failed to identify treatment effect modification in this meta-analysis, although data availability and other issues precluded strong conclusions about heterogeneity. Two recent observational studies based on MS patients followed for 24 and 36 months in two centers in the United States have suggested that after addressing baseline differences in treatment groups, odds of relapses were similar for patients with RRMS treated with DMF and FTY, but the tolerability profile appeared to favor FTY.[7,8]

Using the French MS registry (Observatoire Français de la Sclérose en Plaques [OFSEP]), this study aims to compare the effectiveness of DMF and FTY for preventing relapse recurrence within 12 months in patients with RRMS. Secondary outcomes, including disease progression in terms of disability or radiology and treatment discontinuation, were also examined. We applied an ITR scoring framework to assess whether there were potential subgroups of patients who may benefit more from one of the two treatments.

Methods

The OFSEP cohort

OFSEP (www.ofsep.org) is an ongoing French nationwide registry with longitudinal follow-up of patients diagnosed with MS or who presented with radiologically and clinically isolated syndromes suggestive of MS and MS-related conditions. Data are collected during routine visits (i.e. usually ≥ 1 yearly) using a standardized clinical form. Individual case reports include demographic and MS disease characteristics, clinical and magnetic resonance imaging (MRI) data as well as treatment information. As of June 2018, OFSEP includes 68,000 records from 36 centers, representing over 50% of patients diagnosed with MS nationally in France.[9]

Standard protocol approvals, registrations, and patient consents

All patients enrolled in the OFSEP cohort signed an informed consent form to have their medical data shared anonymously for research purposes. In accordance with the French legislation, the OFSEP cohort was approved by both the national data protection agency (*Commission Nationale de l'Informatique et des Libertés* [CNIL]; authorization request 914066v3) and the French ethical committee (*Comité de Protection des Personnes* [CPP]: reference 2019-A03066-51) and the present study was approved by the Ile-de-France VI Ethics Committee (CNIL 914066v2). For the current study, data were extracted in December 2018.

Patients

Adults with RRMS who initiated DMF or FTY as a second-line treatment between 15 January 2014 and 15 December 2016 were included in this study. RRMS was defined by the 2010 revised MacDonald criteria[10] and the Poser criteria.[11] The definition of a second-line treatment required that the patient had previously stopped the first-line treatment due to inefficacy, experienced relapse(s) following 6 months of first-line treatment or stopped a second-line treatment within 2 years. Inclusion criteria are adults aged at least 18 years old with a valid baseline Expanded Disability Status Scale (EDSS) assessment, defined as the most recent EDSS score 12 weeks before or after treatment initiation and at least 30 days after a relapse. Patients who started DMF or FTY as a first treatment were excluded. Baseline was defined as the date of the first prescription for DMF or FTY as

a second-line treatment. Patients were followed for a minimum of 12 months until the last clinical evaluation or outcome occurrence. Intent-to-treat analyses were considered, that is, patients could be followed beyond treatment (DMF or FTY) discontinuation. Definitions of first- and second-line treatments can be found in Supplemental Methods.

Outcomes and covariates

The primary outcome was annualized relapse rate (ARR) at 12 months. Seven secondary outcomes were defined: time to first relapse, time to treatment discontinuation, the proportion of relapse-free patients at 12 months, the proportion of patients who persisted in treatment at 12 months, the proportion of patients with disability progression at 12 months, the proportion of patients with new T1 gadolinium-enhancing lesions at 12 months, and the proportion of patients with new or newly enlarging T2 lesions at 12 months. Disability progression was defined as a punctual EDSS score increase of $\geq$ 1.0 point for patients with a baseline EDSS score between 0 and 5.5 and an increase of $\geq$ 0.5 point for patients with a baseline EDSS > 5.5, compared to the baseline EDSS score. In sensitivity analyses, the primary and secondary outcomes which are assessed at 12 months were re-evaluated at 24 months.

The following baseline covariates were measured in the 12-month pre-baseline period: number of relapses, EDSS score, disease activity (relapse only, EDSS progression only, relapse and EDSS progression, or none), type of MRI (brain only, spinal only, brain + spinal, brain + optic nerve, brain + spinal + optic nerve), gadolinium contrast enhancement (positive, negative or unknown), number of T2 brain lesions (< 9, ≥ 9 or unknown), and number of T2 spinal lesions (0, 1, ≥ 2, or unknown). Other baseline covariates were age at baseline, sex, MS disease duration in months, number of prior DMTs, and number of relapses in the 24 months pre-baseline.

Statistical methods

General statistical considerations. Continuous and categorical covariates were reported as mean (standard deviation [SD]) or frequency count (proportion), respectively. For treatment group comparisons, *p*-values based on two-sample *t*-tests, Mann-Whitney rank tests, Fisher's exact tests, or Pearson chi-square tests were used as appropriate. We conducted complete case analyses with respect to outcomes and continuous covariates. Missing categorical values were treated as an additional category.

The number of relapses was analyzed with a Poisson regression model, using log-transformed follow-up time as an offset, to estimate ARR ratios. Time-to-event outcomes were analyzed using a Cox proportional hazards model to estimate hazard ratios (HRs). Restricted mean survival time (RMST) differences were also estimated because a doubly robust estimator of the average treatment effect is available for this metric but not for the HR. The RMST is intuitively interpreted as the average event-free survival time up to a pre-specified time point.[12] Censoring for time-to-event outcomes was addressed with inverse probability of censoring weights, which is modeled via fitting Cox proportional hazards models for the censoring time. Binary outcomes (e.g. proportion of relapse-free patients at 12 months) were compared using odds ratios (ORs) estimated from a logistic regression model. *P*-values below 0.05 were considered statistically significant.

Average treatment effect and propensity score. We chose a doubly robust estimator based on propensity score weighting to estimate average treatment effects.[4] The outcome and propensity score models in the doubly robust estimator were adjusted for the following baseline covariates: age at baseline, sex, disease duration, number of prior DMTs, EDSS score, number of relapses in the 12 months pre-baseline, and disease activity in the 12 months pre-baseline. Average treatment effect estimates were also obtained with multivariate-adjusted regression models and a simple unadjusted model. ARR ratios, HRs, and ORs were presented along with 95% confidence intervals, which were obtained via bootstrap with 1000 resamples.

Individualized treatment score and treatment effect heterogeneity. The detection of treatment effect heterogeneity relied on a novel unified precision medicine framework for observational data based on flexible estimation of the ITR score.[4]

The first step requires estimating an ITR score, which is an individual-level DMF versus FTY ARR ratio conditional on the patient's baseline covariates. Thus, the ITR score is a continuous score that represents how much a given individual would benefit from DMF over FTY (or vice-versa) when considering the ARR as the clinical outcome. Here, the ITR score is defined such that lower scores represent larger individual-specific benefits of DMF compared

to FTY. The ITR score can be used to define a subgroup of responders to either DMF or FTY. For example, individuals with the 30% lowest ITR scores are potentially the 30% highest responders to DMF, and individuals with the 30% highest ITR scores are potentially the 30% highest responders to FTY or, equivalently, the 30% standard responders to DMF. In this study, we compared five methods to estimate the ITR score as a function of the following baseline covariates: age at baseline, sex, disease duration, number of prior DMTs, EDSS score, number of relapses in the 12 months pre-baseline, and disease activity in the 12 months pre-baseline.

The second step requires identifying the ITR score method that captured the strongest treatment effect heterogeneity, if present. For this, we use validation curves which provide a visual assessment of treatment effect heterogeneity. For each of the five ITR score methods, the validation curve is constructed by following the steps below:

1. We rank individuals according to their estimated ITR score from the lowest to the highest.
2. We construct a series of nested subgroups of patients with the % lowest scores, for k between 35 and 65. For example, the smallest subgroup constructed with $k = 35$ is composed of the 35% of individuals with the lowest estimated ITR score, which represents the 35% highest responder to DMF, that is, the 35% of individuals who are more likely to respond better to DMF compared to FTY, if treatment heterogeneity is present.
3. We build a validation curve by estimating the average treatment effect with the doubly robust estimator for each subgroup identified in Step 2.

A visual inspection of the validation curve allows a global assessment of treatment effect heterogeneity. For example, a flat validation curve would indicate the absence of treatment effect heterogeneity because the average treatment effect remains unchanged across the subgroups of potential responders to DMF. On the contrary, a steep increasing validation curve would indicate the presence of heterogeneity because the average treatment effect changes across subgroups, meaning that it varies according to the ITR score, and thus according to (a combination of) baseline characteristics. A steeper curve indicates more heterogeneous treatment effects found in the subgroups.

In the last step, we further assess treatment effect heterogeneity with a formal statistical test. For each ITR score method, we split individuals into two subgroups based on the ITR score: potential high responders to DMF and potential standard responders. Then, we use a statistical test based on treatment-by-ITR score interaction models and permutation tests to compare the average treatment effect between the two subgroups. Additional methodological details are available in Supplemental Methods.

Results

Description of the study cohort

The study sample consisted of 1166 patients diagnosed with RRMS, with 73% of female, and the mean age at baseline was 40 years old (SD: 10). A total of 554 patients (48%) were treated with DMF, and 612 (52%) with FTY. Patients on DMF had a lower mean EDSS score (2.1 vs. 2.4, p-value: < 0.001) and had differences in recent disease activity including significantly fewer relapses in the 12 or 24 months prior to baseline (e.g. 1.1 vs. 1.6 in prior 24 months, p-value: < 0.001) compared to those who initiated on FTY (Table 1). Flowcharts summarizing the sample selection procedure are presented in Supplemental Figures 1 and 2.

Number of relapses at 12 months

Average treatment effect. A total of 303 relapses were observed in the 12 months post-baseline, among which 128 relapses occurred in the DMF group and 175 relapses in the FTY group. The unadjusted ARR ratio estimate between DMF and FTY was 0.81 (95% CI: 0.63, 1.04), where an ARR ratio below 1 favor DMF. After weighting, the balance was achieved for all covariates (Table 1 and Supplemental Figure 3). Overall, there was no significant difference between DMF and FTY in terms of 12-month ARR with either the doubly robust estimator (ARR ratio: 1.06, 95% CI: 0.78, 1.43) or the multivariate-adjusted estimator (ARR ratio: 1.02, 95% CI: 0.79, 1.32).

ITR score and treatment effect heterogeneity. The validation curves for the five scoring methods are shown in Figure 1, which allows us to visually assess the presence or absence of treatment effect heterogeneity. The figure shows the ARR ratio DMF versus FTY (i.e. the average treatment effect) in subgroups of high responders to DMF (y-axis) relative to the proportion of patients included in the subgroup (x-axis). Each line corresponds to the validation curve from a different scoring method. All validation curves are relatively flat, meaning that the ITR scores based on the chosen methods are not useful to identify

Table 1. Baseline characteristics of 1166 patients selected from OFSEP by treatment group.

Characteristics	FTY ($n = 612$)	DMF ($n = 554$)	p-value[b]	SMD before weighting	SMD[c] after weighting
Demographics					
Age at baseline	38.9 (10)	40.7 (11)	0.003	0.174	0.003
Male	165 (27)	152 (27)	0.907	0.011	0.003
Disease duration in months	115 (89)	127 (93)	0.020	0.137	0.003
Severity and treatment history					
Number of previous DMTs	2.0 (1.3)	1.8 (1.2)	0.050	0.115	0.005
Number of relapses in the 24 months before baseline	1.6 (1.1)	1.1 (1.1)	< 0.001	0.419	0.020
EDSS score	2.4 (1.6)	2.1 (1.6)	< 0.001	0.220	0.013
Disease activity in the 12 months before baseline					
Relapse only	234 (38)	143 (26)	< 0.001	0.269	0.085
EDSS progression only	18 (3)	27 (5)	0.119	0.100	0.005
Relapse and EDSS progression	50 (8)	23 (4)	0.007	0.168	0.027
None	105 (17)	186 (34)	< 0.001	0.384	< 0.001
Unknown	205 (33)	175 (32)	0.528	0.041	0.068
MRI data in the 12 months before baseline					
MRI[1]			0.413	–	–
Brain only	203 (57)	155 (59)			
Spinal only	17 (5)	6 (2)			
Brain + spinal	128 (36)	99 (37)			
Brain + optic nerve	3 (<1)	1 (< 1)			
Brain + spinal + optic nerve	5 (1)	4 (2)			
Unknown	3 (<1)	0 (0)			
Gadolinium contrast enhancement[a]			0.002	–	–
Positive	147 (41)	75 (28)			
Negative	203 (57)	176 (66)			
Unknown	9 (3)	14 (5)			
Number of brain lesions T2[a]			0.011	–	–
0	2 (< 1)	1 (< 1)			
< 9	27 (8)	26 (10)			
9	267 (79)	176 (68)			
Unknown	43 (13)	56 (22)			
Number of spinal lesions T2[a]			0.378	–	–
0	15 (10)	18 (17)			
1	16 (11)	14 (13)			
2	95 (63)	60 (55)			
Unknown	24 (16)	17 (16)			

DMF: dimethyl fumarate; DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; FTY: fingolimod; MRI: magnetic resonance imaging; SD: standard deviation; SMD: standardized mean (or proportion) difference.

[a]MRI and gadolinium contrast enhancement are available for 359 FTY patients and 265 DMF patients. Brain T2 lesions were measured for 339 FTY patients and 259 DMF patients. Spinal T2 lesions were measured for 150 FTY patients and 109 DMF patients.

[b]P-values are from two-sample t-tests, Mann-Whitney rank tests, Fisher's exact tests, or Pearson chi-square tests.

[c]SMD in the inverse probability of treatment weighted sample, with weights estimated using the propensity score specified as a function of age at baseline, sex, disease duration, number of prior DMTs, EDSS score, number of relapses in the 12 months pre-baseline, and disease activity in the 12 months pre-baseline.

Data are reported as mean (SD) for continuous variables, and n (%) for categorical ones. Long dashes (–) are inserted when the covariates were not considered in the estimation of the overall treatment effect.

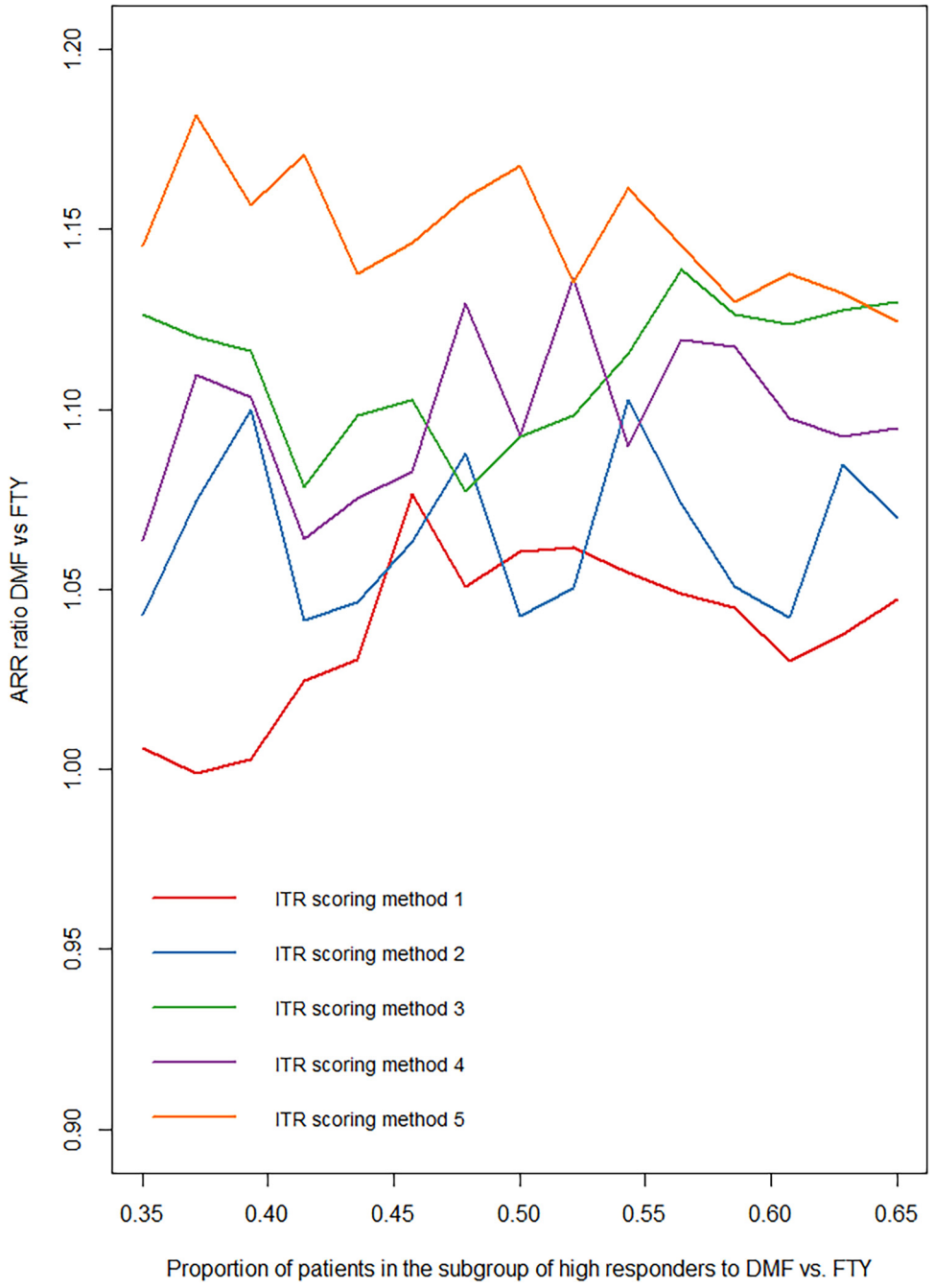

Figure 1. Validation curves for the five scoring methods for the effect of DMF versus FTY on the number of relapses at 12 months after baseline. Flat validation curves, as depicted here, suggest no treatment effect heterogeneity. ARR: annualized relapse rate; DMF: dimethyl fumarate; FTY: fingolimod; ITR: individualized treatment response.

individuals who are better responders to DMF because the average treatment effect in each subgroup remains unchanged. Therefore, we find no evidence of treatment heterogeneity. Visually, we also note that no ITR scoring method evidenced a steeper curve, indicating that no scoring method was superior at detecting treatment heterogeneity.

Table 2 shows the cross-validated ARR ratio DMF versus FTY by subgroups of high responders to DMF versus standard responders for three thresholds k % (35%, 50%, and 65%). For example, using a threshold of 35%, the ARR ratio was 1.00 (95% CI: 0.57, 1.95) for high responders to DMF and 1.06 (95% CI: 0.74, 1.57) for standard responders (p-value for treatment

Table 2. Cross-validated ARR ratios DMF versus FTY by subgroups defined with ITR score 1.

Threshold	Subgroup[a]	ARR ratio DMF versus FTY (95% CI)	*p*-value for heterogeneity
35%	High responders	1.00 (0.57, 1.95)	0.885
	Standard responders	1.06 (0.74, 1.57)	
50%	High responders	1.06 (0.67, 1.67)	0.857
	Standard responders	1.13 (0.68, 1.86)	
65%	High responders	1.05 (0.79, 1.56)	0.913
	Standard responders	1.09 (0.69, 2.18)	

ARR: annualized relapse rate, CI: confidence interval, DMF: dimethyl fumarate, FTY: fingolimod.
[a]For each threshold *k*, the high responder's subgroup includes the *k*% patients with the lowest ITR scores and the standard responder's sugroup includes the 100%-*k*% patients with the highest ITR scores.

heterogeneity: 0.89). Regardless of the choice of threshold, no significant treatment effect heterogeneity was found. We further note that in the absence of treatment effect heterogeneity the subgroup labels of high versus standard responders to DMF become meaningless.

Secondary outcomes

Average treatment effects. Table 3 summarizes the average effect of DMF versus FTY for the seven secondary outcomes. Overall, DMF was significantly associated with a shorter time to treatment discontinuation (doubly robust RMST difference estimate: −0.64, 95% CI: −0.95, −0.34) and lower odds of treatment persistence at 12 months (doubly robust OR estimate: 0.53, 95% CI: 0.39, 0.70) compared to FTY. Reasons for treatment discontinuation are shown in Supplemental Table 3. There was no significant difference between DMF and FTY in terms of time to first relapse and proportion of relapse-free patients, patients with disability progression, patients with new T1 gadolinium-enhancing lesions, or patients with new or newly enlarging T2 lesions at 12 months.

ITR score and treatment effect heterogeneity. We did not explore treatment heterogeneity for the two MRI secondary outcomes due to small sample sizes. Figure 2 shows the validation curves for the five scoring methods for the remaining five secondary outcomes. Most validation curves are flat or present minimal increasing trends (C and E), suggesting little evidence of treatment effect heterogeneity. This was confirmed by a formal test that found no significant treatment effect heterogeneity for any of the secondary outcomes. Additional results and sensitivity analyses on secondary outcomes can be found in Supplemental Tables 4 to 9 and Supplemental Figure 4.

Discussion

A real-world implementation of the personalized scoring approach was conducted to learn average treatment response and to detect and internally validate potential treatment effect heterogeneity of DMF versus FTY among RRMS patients in the OFSEP registry. Findings suggest no difference in average effectiveness of DMF versus FTY for preventing relapse recurrence when used as a second-line therapy. Disability progression and MRI-based disease activity outcomes were also similar between treatments. We observed a lower proportion of patients who persisted on treatment at 12 months for DMF versus FTY. These findings are consistent with a growing body of literature from observational studies that have demonstrated similar effectiveness of these DMTs for preventing relapses and suggested a better tolerability profile and less discontinuation for patients treated with FTY.[7,8,13–19]

The five ITR scoring methods did not identify significant treatment effect heterogeneity with respect to any primary or secondary outcomes, implying that not enough evidence was gathered to conclude that the effect of DMF versus FTY varies across patients in terms of the studied outcomes. While the ITR scoring methods are not accompanied by formal power calculation, their statistical robustness and ability to handle several treatment effect modifiers simultaneously make the scoring methods more attractive than traditional subgroup analyses, thus providing reassurance on the robustness of the results. Nevertheless, a comparison of DMF versus FTY as second-line treatment in other real-world data sources, ideally with larger sample sizes, remains warranted to provide stronger evidence on the absence of treatment effect heterogeneity.

The precision medicine framework applied in the present observational study is a relatively novel approach to characterizing individualized treatment

Table 3. Unadjusted, multivariate-adjusted, and doubly robust estimators of the overall effect of DMF versus FTY for primary and secondary outcomes assessed at 12 months.

		Unadjusted	Multivariate-adjusted	Doubly robust[d]
Primary and secondary outcomes	*n*	HR/OR (95% CI)	HR/OR (95% CI)	RMST difference/OR (95% CI)
Time to event outcomes				
Time to first relapse	1166	0.84 (0.70, 1.01)	1.08 (0.89, 1.30)	−0.25 (−0.57, 0.06)
Time to treatment discontinuation	1166	1.46 (1.23, 1.74)	1.74 (1.45, 2.09)	−0.64 (−0.95, −0.34)
Proportions at 12 months				
Relapse-free patients	1166	1.31 (0.98, 1.74)	0.99 (0.73, 1.35)	0.98 (0.75, 1.28)
Patients who persisted on treatment	1166	0.62 (0.47, 0.82)	0.53 (0.40, 0.71)	0.53 (0.39, 0.70)
Patients with disability progression[a]	833	0.78 (0.54, 1.13)	0.78 (0.53, 1.16)	0.76 (0.54, 1.06)
Patients with new T1 GdE lesions[b]	428	0.66 (0.36, 1.21)	0.69 (0.37, 1.29)	0.67 (0.37, 1.23)
Patients with new or newly enlarging T2 lesions[c]	447	0.66 (0.43, 1.01)	0.73 (0.46, 1.14)	0.73 (0.50, 1.07)

CI: confidence interval; DMF: dimethyl fumarate; FTY: fingolimod; GdE: gadolinium-enhancing; HR: hazard ratio DMF versus FTY; OR: odds ratio DMF versus FTY; RMST: restricted mean survival time (difference DMF minus FTY). HR and OR have FTY as a reference category. RMST difference is expressed as DMF minus FTY. For the time-to-event outcomes, HR < 1 and RMST > 1 favor DMF. For proportions at 12 months, OR < 1 favors DMF for all outcomes except for patients who persisted in treatment for which OR < 1 favors FTY.
The multivariate-adjusted and doubly robust models were adjusted for the following covariates: age at baseline, sex, disease duration, number of prior DMTs, EDSS score, number of relapses in the 12 months pre-baseline, and disease activity in the 12 months pre-baseline.
[a]378 DMF, 455 FTY.
[b]179 DMF, 249 FTY. Of the 447 patients available for the analysis related to T1 GdE lesions, 19 patients were excluded because they did not have information on the status of gadolinium. For T1 GdE lesions, propensity score and outcome models were not adjusted for baseline T1 GdE lesions.
[c]187 DMF, 260 FTY. For new or newly enlarging T2 lesions, propensity score and outcome models were not adjusted for baseline new or newly enlarging T2 lesions.
[d]For the time-to-event outcomes, a doubly robust estimator of the HR cannot be calculated. Instead, RMST differences are reported, where RMST differences below 0 favor FTY over DMF for both time-to-event outcomes.

responses. The completeness of the approach (ITR score, validation curves, formal statistical test) suggests that it is a promising method to apply for future real-world analyses wherein treatment effect modification is suspected. The framework can be used to either confirm that the average treatment effect is an appropriate analytical approach, as in the present study or to prompt the development of enrichment strategies and further investigation into subgroups that might benefit more from a treatment than the average patient.

This study is subject to limitations. First, while doubly robust estimators based on propensity scores were used to address imbalances in covariates by treatment arm, both overall and within subgroups, there remains the potential for residual unmeasured confounding. For example, baseline MRI data are likely important

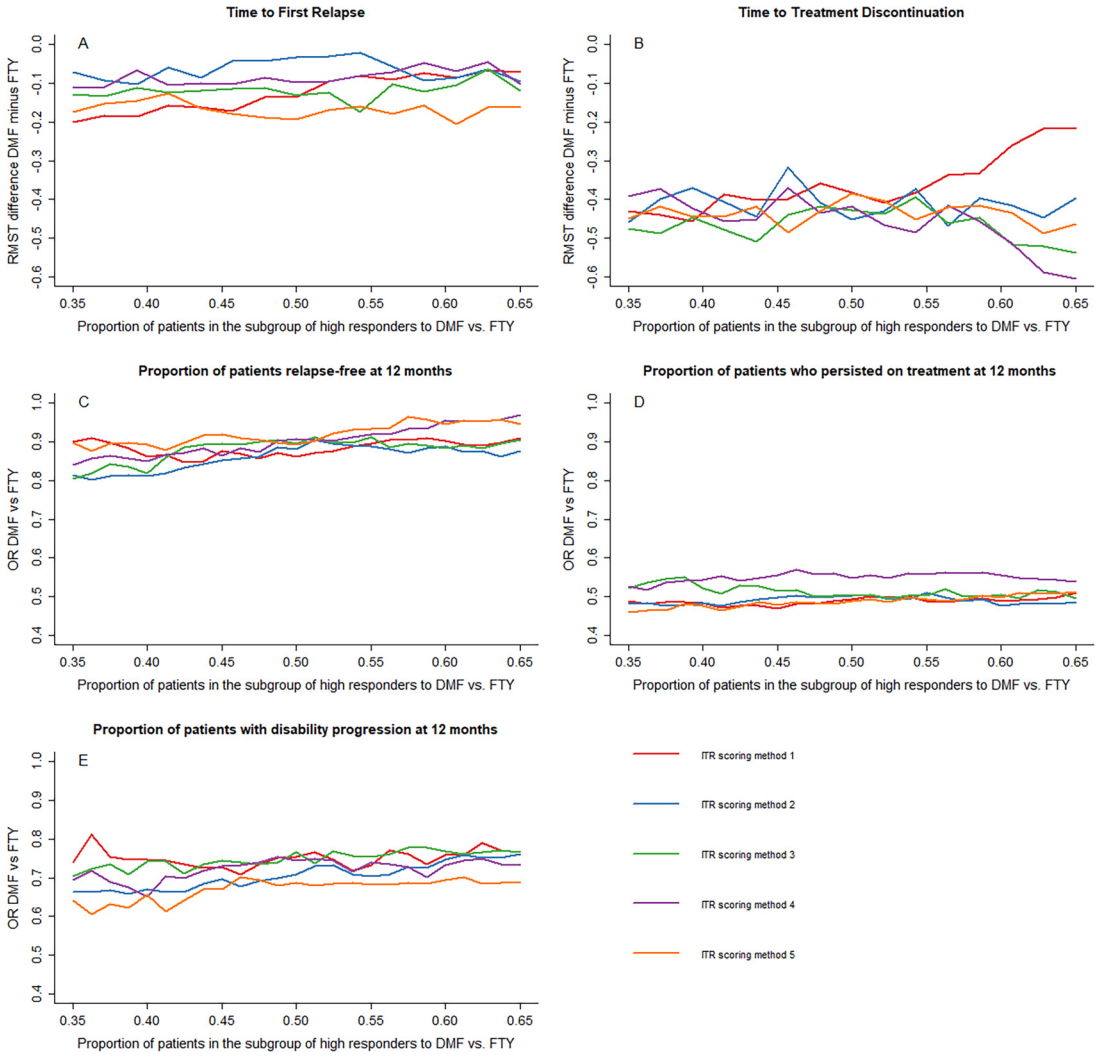

Figure 2. Validation curves for the five scoring methods for the effect of DMF versus FTY on time to first relapse (A), time to treatment discontinuation (B), the proportion of relapse-free patients at 12 months (C), the proportion of patients who persisted in treatment at 12 months (D) and proportion of patients with disability progression at 12 months (E). Flat validation curves, as depicted here, suggest no treatment effect heterogeneity. DMF: dimethyl fumarate; FTY: fingolimod; ITR: individualized treatment response; OR: odds ratio; RMST: restricted mean survival time.

confounders in the treatment–outcome relationships considered, yet they were not considered in the analysis due to their unavailability for a substantial portion of our cohort. Also, by using LASSO penalization to fit the PS model, we may have inadvertently omitted weak confounders while selecting strong predictors of treatment assignment, which goes against existing guidance for covariate selection in the PS.[20] Nevertheless, covariate balance was satisfactory after PS weighting. Second, while OFSEP is a rich data source with high coverage of the target population in France, missing data may have impacted the findings due to exclusion of patients with missing EDSS at baseline. Similarly, the timing of the routine visits, which was not standardized, could be associated with the clinical outcomes, leading to informative visit patterns which we did not address.[21] Third, while sample selection criteria aimed toward selecting a well-defined population initiating DMF or FTY as a second-line therapy for RRMS, they may limit the generalizability of findings. Moreover, these restrictions on the study population may have contributed to the absence of significant treatment heterogeneity observed by limiting patient heterogeneity. Lastly, cross-validation was used to assess the performance of each ITR scoring method, which may overestimate performance as compared to validation performed using external data. External validation would be a natural future direction for this analysis.

This study provided the first formal precision medicine exploration of DMF versus FTY using a novel and robust statistical approach based on a continuous scoring system. Personalized approaches to the treatment of MS remain sparse, and this application sets the ground for future head-to-head comparisons of DMTs in real-world settings.

Acknowledgements

Precision Analytics Inc (Montreal, QC, Canada) participated in manuscript preparation.

Data availability

Data that support the findings of this study are available upon a motivated request to the OFSEP coordinator and will be evaluated by the study coordinator and the OFSEP scientific committee for approval according to OFSEP bylaws and access to data procedures.

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: FP, MR, CM, GS, XJ, and CdM are employees of Biogen and hold stocks of the company. LT and MC received consulting fees from Biogen. The funder participated in the study conception and design, interpretation of findings, and drafting of the manuscript. FP had full access to all of the data in this study and takes responsibility for the integrity of the data and the accuracy of the data analysis. SV and FR have nothing to disclose.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Biogen. This work was conducted independently of the funding source using data from the *Observatoire Français de la Sclérose en Plaques* (OFSEP) which is supported by a grant provided by the French State and handled by the *Agence Nationale de la Recherche*, within the framework of the Investments for the Future program, under the reference ANR-10-COHO-002 and by the Eugène Devic EDMUS Foundation against multiple sclerosis.

ORCID iDs

Gabrielle Simoneau https://orcid.org/0000-0001-9310-6274
Massimiliano Copetti https://orcid.org/0000-0002-7960-5947
Sandra Vukusic https://orcid.org/0000-0001-7337-7122

Supplemental material

Supplemental material for this article is available online.

References

1. Tramacere I, Del Giovane C, Salanti G, et al. Immunomodulators and immunosuppressants for relapsing-remitting multiple sclerosis: a network meta-analysis. *Cochrane Database Syst Rev*. 2015. CD011381.
2. Rotstein D and Montalban X. Reaching an evidence-based prognosis for personalized treatment of multiple sclerosis. *Nat Rev Neurol*. 2019; 15: 287–300.
3. Pellegrini F, Copetti M, Bovis F, et al. A proof-of-concept application of a novel scoring approach for personalized medicine in multiple sclerosis. *Mult Scler J*. 2019; 26(3): 1064–1073.
4. Yadlowsky S, Pellegrini F, Lionetto F, et al. Estimation and validation of ratio-based conditional average treatment effects using observational data. *J Am Stat Assoc*. 2020(): 116(533): 335–352.
5. Brookes ST, Whitely E, Egger M, et al. Subgroup analyses in randomized trials: risks of subgroup-specific analyses: power and sample size for the interaction test *J Clin Epidemiol*. 2004;57: 229–236.
6. Li J, Zhao L, Tian L, et al. A predictive enrichment procedure to identify potential responders to a new therapy for randomized, comparative controlled clinical studies. *Biometrics*. 2016;72: 877–887.
7. Vollmer B, Ontaneda D, Bandyopadhyay A, et al. Discontinuation and comparative effectiveness of dimethyl fumarate and fingolimod in 2 centers. *Neurol: Clin Pract*. 2018;8: 292–301.
8. Vollmer B, Ontaneda D, Harris H, et al. Comparative discontinuation, effectiveness, and switching practices of dimethyl fumarate and fingolimod at 36-month follow-up. *J Neurol Sci*. 2019;407: 116498.
9. Vukusic S, Casey R, Rollot F, et al. Observatoire français de la sclérose en plaques (OFSEP): a unique multimodal nationwide MS registry in France. *Mult Scler J*. 2020;26: 118–122.
10. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol*. 2011;69: 292–302.
11. Poser CM, Paty DW, Scheinberg L, et al. New diagnostic criteria for multiple sclerosis: guidelines for research protocols. *Ann Neurol*. 1983;13: 227–231.
12. Kim DH, Uno H and Wei L-J. Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol*. 2017;2: 1179–1180.
13. Wicks P, Rasouliyan L, Katic B, et al. The real-world patient experience of fingolimod and dimethyl fumarate for multiple sclerosis. *BMC Res Notes*. 2016;9: 1–9.
14. Hersh CM, Love TE, Cohn S, et al. Comparative efficacy and discontinuation of dimethyl fumarate and fingolimod in clinical practice at 12-month follow-up. *Mult Scler Relat Disord*. 2016;10: 44–52.
15. Hersh CM, Love TE, Bandyopadhyay A, et al. Comparative efficacy and discontinuation of dimethyl fumarate and fingolimod in clinical practice at 24-month follow-up. *Mult Scler J–Exp, Transl Clin*. 2017;3: 2055217317715485.

16. Vollmer B, Nair KV, Sillau SH, et al. Comparison of fingolimod and dimethyl fumarate in the treatment of multiple sclerosis: two-year experience. *Mult Scler J–Exp, Transl Clin*. 2017;3: 2055217317725102.
17. Prosperini L, Lucchini M, Haggiag S, et al. Fingolimod vs dimethyl fumarate in multiple sclerosis: a real-world propensity score-matched study. *Neurology*. 2018;91: e153–ee61.
18. Braune S, Grimm S, van Hövell P, et al. Comparative effectiveness of delayed-release dimethyl fumarate versus interferon, glatiramer acetate, teriflunomide, or fingolimod: results from the German NeuroTransData registry. *J Neurol*. 2018;265: 2980–2992.
19. Kalincik T, Havrdova EK, Horakova D, et al. Comparison of fingolimod, dimethyl fumarate and teriflunomide for multiple sclerosis. *J Neurol Neurosurg Psychiatry*. 2019;90: 458–468.
20. Ali MS, Groenwold RH and Klungel OH. Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. *Am J Clin Nutr*. 2016;104: 247–258.
21. Pullenayegum EM and Lim LS. Longitudinal data subject to irregular observation: a review of methods with a focus on visit processes, assumptions, and study design. *Stat Methods Med Res*. 2016;25: 2992–3014.