MSJ Experimental Translational Clinical

Original Research Article

Overall and patient-specific comparative effectiveness of dimethyl fumarate versus teriflunomide: A novel approach to precision medicine applied to the German NeuroTrans Data Multiple Sclerosis Registry

Multiple Sclerosis Journal—Experimental, Translational and Clinical

July–September 2023, 1–10

DOI: 10.1177/20552173231194353

Xiaotong Jiang, Gabrielle Simoneau[#], Mel Zuercher, Yanic Heer, Philip van Hoevell, Adrian Harrington, Wanda Castro-Borrero, Carl de Moor[#], Fabio Pellegrini, Lu Tian, Arnfin Bergmann and Stefan Braune

Abstract

Background: Multiple sclerosis (MS) comparative effectiveness research needs to go beyond average treatment effects (ATEs) and post-host subgroup analyses.

Objective: This retrospective study assessed overall and patient-specific effects of dimethyl fumarate (DMF) versus teriflunomide (TERI) in patients with relapsing-remitting MS.

Methods: A novel precision medicine (PM) scoring approach leverages advanced machine learning methods and adjusts for imbalances in baseline characteristics between patients receiving different treatments. Using the German NeuroTransData registry, we implemented and internally validated different scoring systems to distinguish patient-specific effects of DMF relative to TERI based on annualized relapse rates, time to first relapse, and time to confirmed disease progression.

Results: Among 2791 patients, there was superior ATE of DMF versus TERI for the two relapse-related endpoints ($p = 0.037$ and 0.018). Low to moderate signals of treatment effect heterogeneity were detected according to individualized scores. A MS patient subgroup was identified for whom DMF was more effective than TERI ($p = 0.013$): older (45 versus 38 years), longer MS duration (110 versus 50 months), not newly diagnosed (74% versus 40%), and no prior glatiramer acetate usage (35% versus 5%).

Conclusion: The implemented approach can disentangle prognostic differences from treatment effect heterogeneity and provide unbiased patient-specific profiling of comparative effectiveness based on real-world data.

Keywords: Dimethyl fumarate, teriflunomide, comparative effectiveness, treatment effect heterogeneity, propensity score, precision medicine, real-world data

Date received: 27 March 2023; accepted 24 July 2023

Introduction

Multiple sclerosis (MS) is one of the most prevalent neurological autoimmune diseases and may cause severe disability. More than a dozen disease-modifying therapies (DMTs) are approved for reducing the risk of MS relapses, such as dimethyl fumarate (DMF) and teriflunomide (TERI), but there is currently no single solution that controls disease progression for all patients with MS. MS is heterogeneous in both disease course and patient responses, motivating the development of a personalized strategy. It is important to prescribe the best treatment for each individual to optimize the timing of treatment, better manage symptoms, and prevent future relapses and loss in functions and quality of life. To help clinicians make informed treatment decisions and to save patients from unwarranted physical and financial burden, precision medicine using rich medical data of patients with MS can serve as a

[#]Former Biogen employee.

Correspondence to: Xiaotong Jiang, Biogen, 225 Binney Street, Cambridge, MA 02142, USA. Email: phoebe.jiang@biogen.com

Xiaotong Jiang, Biogen, Cambridge, MA, USA

Gabrielle Simoneau, Biogen Canada, Toronto, Ontario, Canada

Mel Zuercher, Yanic Heer, Philip van Hoevell, Rewoso AG, Zürich, Switzerland

Adrian Harrington, Biogen International GmbH, Baar, Switzerland

Wanda Castro-Borrero, Biogen International GmbH, Baar, Switzerland

Wanda Castro-Borrero, Biogen International GmbH, Baar, Switzerland

Wanda Castro-Borrero, Biogen International GmbH, Baar, Switzerland

Wanda Castro-Borrero, Carl de Moor, Biogen, Cambridge, MA, USA

Fabio Pellegrini, Biogen Spain, Madrid, Spain

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

Arnfin Bergmann, Stefan Braune, NeuroTransData, Neuburg an der Donau, Germany

promising guide. Precision medicine is a data-driven paradigm that recommends the most effective treatment for individual patients based on patient-specific clinical variables (i.e., treatment effect modifiers).

The objective of this analysis is to explore the potential treatment effect heterogeneity between two oral DMTs, DMF and TERI, in the German NeuroTransData (NTD) registry and to look for stratifications of the patient population for tailored treatment recommendation. Real-world evidence has several advantages over evidence derived from randomized controlled trials (RCTs), which are often underpowered for detecting patient-specific heterogeneity[1]: (1) acquisition of larger sample sizes; (2) more variety of patient characteristics; and (3) treatment administration representative of the real-world setting. A recently developed generalization of a novel precision medicine method used for RCTs has been applied (Supplemental Figure S1).[2,3] The method stratifies patients with similar characteristics based on a scoring system, which measures the relative benefit of DMF versus TERI for each individual based on their baseline characteristics. The method is especially appealing in analyzing treatment effect for patients with MS because it: (1) adjusts for potential confounding; (2) constructs and internally validates the scoring systems; and (3) caters to ratio-based metrics of the treatment benefit, suitable for count and survival outcomes. This work extends the previous analysis of two phase 3 RCTs between DMF versus placebo,[1] a clinical application of the original methodological work[2,3] tailored to a clinical audience. It is also a companion analysis of a French MS registry analysis for comparative treatment effectiveness between DMF and fingolimod.[4] By the application of this precision medicine approach to real-world data, we aim to provide guidance on the personalized selection between two MS therapeutic options and pave the road to personalized treatment decisions in MS.

Methods

The NTD network and NTD MS registry

Founded in 2008, the NTD is a physician network of approximately 66 practices and 133 members across Germany in the field of neurology and psychiatry, gathering healthcare data to improve physician–patient interaction and optimize the treatment of individual patients.[5] Information on demographics, medical history, patient-related outcomes, and clinical variables are recorded regularly via a web-based patient management platform[6] during outpatient office visits, with 3.7 visits per year on average.[7] A comprehensive manual and automated data control framework has been established to ensure high data quality.[8] As of the end of 2019, around 25,000 patients with MS were included and followed longitudinally over an average of 5 years in the NTD registry, representing around 10% of approximately 224,000 patients with MS in Germany.[9]

Study population

The study included NTD records of 2791 patients with relapsing-remitting MS who received DMF or TERI between January 1, 2009, and July 1, 2018. Inclusion and exclusion criteria have been published.[7] Patient baseline characteristics were taken at the initiation of index therapy.

Effectiveness assessment endpoints

Annualized relapse rate (ARR), the average number of relapses[a] for a group of patients per year, was the primary endpoint of this analysis. Secondary endpoints were time to first relapse since treatment initiation and time to 12-week confirmed disease progression (CDP) defined by Expanded Disability Status Scale (EDSS) scores.[b] The time to first relapse or time to CDP may be right censored, and the censoring time was defined as time to loss to follow-up, treatment discontinuation, or data cut date, whichever occurred first.

Baseline characteristics

Nineteen baseline characteristic variables were included both as potential confounders and potential treatment effect modifiers: age, sex, number of prior DMTs, number of months since MS diagnosis, newly diagnosed (yes if diagnosed <12 months ago or no prior DMTs; no otherwise), prior use of glatiramer acetate (GA) (yes/no), prior use of interferon (yes/no), number of relapses in 12 months and in 24 months before treatment initiation, EDSS total score, individual EDSS score (visual, ambulatory, brainstem, cerebellar, cerebral, pyramidal, sensory, sphincteric), and EuroQol-5D-5L visual analog scale (EQ5D5L VAS) score.

Statistical methods

General statistical considerations. Baseline patient characteristics were reported as mean (SD) or *n* (%) for continuous and categorical variables, respectively. Cohen's standardized mean or proportion differences (SMD) were reported to measure the difference in baseline characteristics between treatment groups. Absolute values of SMD > 0.1 were considered clinically relevant. Missing values were imputed by the corresponding observed sample mean.

Baseline variable selection. Three sets of potential confounders, one set for each endpoint, were identified from baseline variables that were associated with the

endpoint at the significance level of 0.1 in a multivariable regression analysis, separately among patients receiving DMF or TERI. Negative binomial regression was used for the ARR, and Cox proportional hazards model was used for the two time-to-event endpoints. The selected informative variables were treated as potential confounders and effect modifiers in the calculation of average treatment effects (ATEs) as well as individualized treatment response (ITR) score.

Average treatment effects. The ATE between DMF and TERI was measured as an ARR ratio for number of relapses and restricted mean time lost (RMTL) ratio for the two time-to-event endpoints. To account for potential confounding, we estimated the ATE adjusting for baseline variables with a doubly robust procedure, which combined the propensity score (PS) model and the outcome regression model for protection against model misspecification and provided more precise statistical inferences on treatment effects.[3] Patients were weighted based on an inverse probability weighting scheme estimated from the PS model to balance their baseline characteristics between the two treatment groups (pages 3–4 in the Supplement). Both unadjusted and doubly robust estimators, together with their 95% confidence intervals, were reported with standard errors obtained from 400 bootstrap samples.

Individualized treatment response score and treatment effect heterogeneity assessment. Patient-specific treatment effects between DMF and TERI were measured with ITR scores, which were estimated by the aforementioned novel precision medicine approach[3] chosen for its robustness and adaptability to real-world data.

Constructed for each of the three endpoints separately, the estimated ITR score for patient i was a ratio between DMF versus TERI,

$$\widehat{\mathrm{ITR}}(x_i) = \frac{\hat{R}^{\mathrm{DMF}}(x_i)}{\hat{R}^{\mathrm{TERI}}(x_i)},$$

where $\hat{R}^{\mathrm{DMF}}(x)$ and $\hat{R}^{\mathrm{TERI}}(x)$ represented the estimated ARR or RMTL of a patient with the baseline variable x_i assuming this patient received DMF and TERI, respectively. The ITR scores were estimated from four scoring methods, ranging from a Poisson regression to more complex machine learning models (e.g., boosting and doubly robust regression methods). A lower ITR score would imply that the methods found more benefits of DMF relative to TERI and vice versa.

We used cross-validated validation curves to evaluate the performance of ITR scores in detecting heterogeneous treatment effects.[3] Validation curves < 1 mean that patients on average benefited more from DMF than TERI, and patients benefited more from TERI on average if >1. The steeper the validation curves, the stronger the treatment effect heterogeneity captured by the ITR scores. To test whether the treatment effects between DMF and TERI varied among patients grouped according to ITR scores, we sorted and grouped patients in the validation set according to their ranked ITR scores into two subgroups, high DMF responders and equal responders, defined a priori based on the 60% versus 40% ITR distributions. We compared the univariate patient baseline data within subgroups and by treatment group to study the characterization of the responder groups. In addition, three mutually exclusive subgroups of approximately equal sizes were identified and the ATE in each subgroup was estimated, the distribution of which were summarized as boxplots across 200 training and validation splits. Two-sided p values < 0.05 were considered statistically significant throughout the study. All analyses were conducted using R version 4.0.5.[10] See Technical Details on Methods section of the Supplement for more technical details.

Results

The analysis included 2791 NTD patients with MS: 1741 (62%) received DMF and 1050 (38%) received TERI. The two treatment cohorts were moderately different. Younger patients with shorter disease duration, more relapses in the prior 24 months, higher EQ5D5L VAS scores, and lower EDSS pyramidal, sensory, and total scores at baseline tended to be treated with DMF (Table 1). This NTD cohort was generally older, had prior DMTs, and had longer MS duration than clinical trial cohorts, as the latter typically enroll patients with fewer complicating characteristics.

Number of relapses at 12 months

Average treatment effect. Average follow-up time was 2.17 (SD 1.73) in the DMF group and 2.11 (SD 1.72) years in the TERI group. DMF patients had a lower unadjusted ARR than TERI patients (DMF 0.33, TERI 0.41). The ARR of patients receiving DMF was 23.7% lower than that of patients receiving TERI (Table 2). Nine baseline variables were identified and balanced after PS weighting (Supplemental Tables S1 and S2). The doubly robust adjusted ARR ratio remained close to the unadjusted ARR ratio (Table 2), suggesting that the ARR among DMF patients was

Table 1. Patient baseline characteristics by treatment group.

Baseline characteristics	DMF *n* (%) 1741 (62.4)	TERI *n* (%) 1050 (37.6)	SMD[a]
Female	1270 (73.0)	732 (69.7)	0.073
Age, years	**39.9 (10.7)**	**44.9 (10.2)**	**0.473**
Prior DMTs, No.	0.96 (0.98)	0.97 (0.93)	0.006
Time since MS diagnosis, months	**78.8 (79.2)**	**97.3 (91.7)**	**0.216**
Relapses in prior 12 months, No.	0.46 (0.65)	0.42 (0.60)	0.064
Relapses in prior 24 months, No.	**0.71 (0.90)**	**0.64 (0.84)**	**0.077**
New diagnosis	698 (40.1)	400 (38.1)	0.041
Prior GA (yes/no)	1327 (76.2)	821 (78.2)	0.048
Prior interferon (yes/no)	886 (50.9)	502 (47.8)	0.062
EQ5D5L VAS score	**86.4 (12.8)**	**84.6 (13.7)**	**0.132**
EDSS total score	**1.84 (1.50)**	**2.03 (1.51)**	**0.127**
EDSS visual score	0.28 (0.64)	0.31 (0.64)	0.048
EDSS ambulatory score	0.23 (0.92)	0.26 (1.06)	0.036
EDSS brainstem score	0.20 (0.49)	0.23 (0.52)	0.061
EDSS cerebellar score	0.48 (0.85)	0.46 (0.81)	0.023
EDSS cerebral score	0.47 (0.75)	0.49 (0.76)	0.028
EDSS pyramidal score	**0.77 (1.04)**	**0.92 (1.10)**	**0.146**
EDSS sensory score	**0.76 (0.88)**	**0.85 (0.91)**	**0.104**
EDSS sphincteric score	0.27 (0.61)	0.27 (0.58)	0.002

DMF: dimethyl fumarate; DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; EQ5D5L: EuroQol-5 Dimension 5-level version; GA: glatiramer acetate; SD: standard deviation; SMD: standardized mean or proportion difference; TERI: teriflunomide; VAS: visual analogue scale.
Data are reported as mean (SD) for continuous variables, and *n* (%) for categorical variables.
New diagnosis is defined as "yes" if one of the following conditions is satisfied: (1) having a time since diagnosis of < 12 months or (2) having no prior DMTs; and "no" otherwise.
[a]Standardized mean or proportion difference (Cohen's d values): In general, a value <0.2 is considered acceptable, between 0.2 and 0.5 considered as a moderate difference, between 0.5 and 0.8 as a significant difference, and >0.8 as a major difference.

significantly reduced compared with TERI patients even after adjusting for confounding effects.

Performance of the ITR score and treatment effect heterogeneity. Figure 1 (top left) displays the validation curves as the performance of the four ITR scoring methods. For each scoring method (colored line), a proportion of patients with the lowest estimated ITR scores was grouped together (*X*-axis) and the observed ARR ratio between DMF and TERI (*Y*-axis) of the subgroup was averaged over 200 cross validationss. Subgroups of patients with smaller proportions of the lowest estimated ITR scores (left part of the *X*-axis) tended to have larger observed treatment effect between DMF and TERI (ARR ratios further from 1). All methods except the boosting method (black) indicated treatment heterogeneity because they produced positive, steeper curves. Two regression and contrast regression (green and blue) were relatively better at detecting treatment heterogeneity than Poisson regression (red). Boosting started to detect treatment heterogeneity when subgroup sizes were larger than 80% of all samples. More explorations of ARR ratios and subgroups can be found on pages 8–11 of the Supplement.

We tested whether the treatment effect between DMF and TERI varied among patients grouped according to ITR scores using the 60% and 40% proportion cut-off (Table 2). Among patients with the lowest 60% ITR scores estimated by contrast regression, which we referred to as *high DMF responders*, the average cross-validated ARR ratio was well below 1, suggesting that DMF was associated with substantially reduced relapse rates compared with TERI in this group. For the remaining patients, which were referred to as *equal responders* to DMF versus TERI, our methods did not provide specific recommendation on treatment selection. The relative benefit of DMF was greater

Table 2. Average treatment effect and treatment effect by subgroups for all three endpoints.

Average treatment effect				
Endpoint	**Model**	**Ratio [DMF versus TERI]**	**95% CI**	***p*-Value**
ARR ratio	Unadjusted	0.76	(0.65–0.90)	0.001
	Doubly robust adjusted[a]	0.78	(0.61–0.98)	0.037
RMTL ratio of time to first relapse	Unadjusted	0.82	(0.72–0.92)	0.001
	Doubly robust adjusted[b]	0.74	(0.58–0.95)	0.018
RMTL ratio of time to CDP	Unadjusted	0.93	(0.78–1.10)	0.368
	Doubly robust adjusted[c]	0.95	(0.70–1.28)	0.717
Treatment effect by subgroups[d]				
				***p*-Value**
Endpoint	**Subgroups**	**Cross-validated ratio [DMF vs. TERI]**	**95% CI**	**Within subgroup** / **Between subgroups**
ARR ratio	**High DMF responders (60%)**	**0.62**	**(0.44–0.88)**	**0.008** / 0.119
	Equal responders (40%)	0.96	(0.67–1.38)	0.832
RMTL ratio of time to first relapse	**High DMF responders (60%)**	**0.60**	**(0.44–0.83)**	**0.002** / 0.004
	Equal responders (40%)	1.22	(0.89–1.68)	0.221
RMTL ratio of time to CDP	High DMF responders (60%)	0.87	(0.66–1.14)	0.300 / 0.686
	Equal responders (40%)	1.05	(0.46–2.43)	0.903

ARR: annualized relapse rate; CDP: confirmed disease progression; CI: confidence interval; DMF: dimethyl fumarate; DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; EQ5D5L: EuroQol-5 Dimension 5-level version; GA: glatiramer acetate; ITR: individualized treatment response; RMTL: restricted mean time lost; TERI: teriflunomide; VAS: visual analogue scale.

p-Values < 0.05 are boldfaced, indicating that ARR ratio and RMTL ratio of time to first relapse are significantly different within the high DMF responder subgroup.

[a]Adjusted for age, number of prior DMTs, months since diagnosis, number of relapses in the previous 24 months, prior GA, prior interferon, EDSS total score, EDSS cerebral score, and EDSS pyramidal score.

[b]Adjusted for age, number of prior DMTs, months since diagnosis, number of relapses in the previous 12 months, number of relapses in the previous 24 months, prior GA, prior interferon, EDSS cerebellar score, and EDSS cerebral score.

[c]Adjusted for number of prior DMTs, months since diagnosis, prior interferon, EQ5D5L VAS score, EDSS total score, EDSS ambulatory score, EDSS cerebellar score, EDSS sensory score, and EDSS sphincteric score.

[d]Based on cross-validated ITR score estimated from contrast regression.

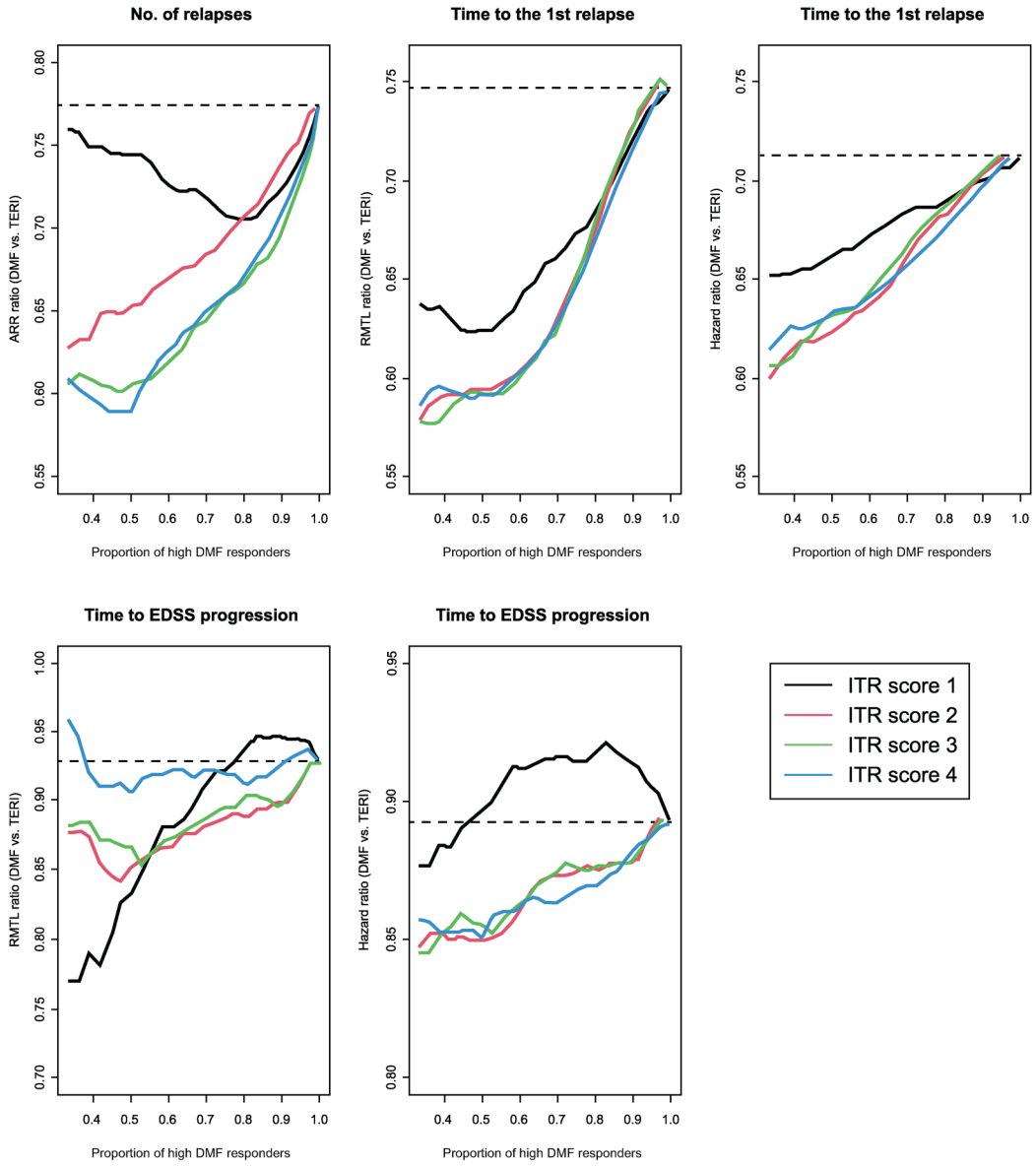

Figure 1. Observed ARR ratio of DMF/TERI (a surrogate for the treatment effect) in nested subgroups of patients ranked by increasing values of the estimated ITR score (high DMF responders), averaged over all validation sets in cross validation. The size of the subgroup of high DMF responders (proportion of patients with the lowest estimated ITR scores): *X*-axis; the observed ARR ratio of DMF/TERI (treatment effect of DMF relative to TERI): *Y*-axis. ITR score 1, boosting (black); ITR score 2, Poisson regression (red); ITR score 3, two regressions (green); ITR score 4, contrast regression (blue). ARR: annualized relapse rate; DMF: dimethyl fumarate; EDSS: Expanded Disability Status Scale; ITR: individualized treatment rule; TERI: teriflunomide.

among the high DMF responders than the equal responders, showing that ITR score possibly detected treatment effect heterogeneity in the ARR ratios. More evidence is needed to conclude that the treatment heterogeneity between the two subgroups was statistically significant ($p = 0.12$). Other proportion cut-offs (33%–67%) were also explored and the observed differences in treatment effects for these subgroups were similar and summarized in Supplemental Table S3.

The weights of the baseline characteristics estimated by the contrast regression scoring method were listed in Table 3, and patient baseline characteristics were compared between the high DMF responder and equal responder subgroups in Table 4. High DMF responders tended to be older (45 versus 38 years of age), had more prior DMTs (1.2 versus 0.65), had a longer MS duration (110 versus 50 months), were not newly diagnosed

Table 3. Weights of the ITR score based on contrast regression for ARR between DMF and TERI.

Baseline characteristics	Weight (95% CI[a])	Standardized weight[b]
Age, years	−0.020 (−0.047, 0.006)	−0.220
Prior DMTs, No.	0.086 (−0.292, 0.463)	0.082
Time since diagnosis, months	−0.003 (−0.008, 0.002)	−0.240
Prior GA (yes/no)	0.722 (0.053, 1.391)	0.304
Prior interferon (yes/no)	0.319 (−0.325, 0.963)	0.159
Relapses in the previous 24 months, No.	0.042 (−0.176, 0.260)	0.037
EDSS total score	0.188 (−0.037, 0.413)	0.283
EDSS cerebral score	0.304 (−0.083, 0.691)	0.229
EDSS pyramidal score	−0.033 (−0.353, 0.287)	−0.035

ARR: annualized relapse rate; CI: confidence interval; DMF: dimethyl fumarate; DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; GA: glatiramer acetate; ITR: individualized treatment response; TERI: teriflunomide.

[a]95% CI and p value were calculated assuming that the true log ARR ratio was a linear combination of baseline variables.

[b]The standardized weights were for the rescaled predictors with a standard deviation of 1 and used to gauge the relative importance of the predictor in the construction of the ITR score. ITR Score = −0.592 − 0.020 × age + 0.086 × number of prior DMTs − 0.003 × months since diagnosis + 0.722 × prior GA + 0.319 × prior interferon + 0.042 × number of relapses in the previous 24 months + 0.188 × EDSS score + 0.304 × EDSS cerebral score–0.033 × EDSS pyramidal score.

(74% versus 40%), had no prior GA (35% versus 5%), and had lower EDSS cerebral scores (0.29 versus 0.72). Further comparison of patient characteristics between DMF and TERI within each responder group can be found in the Supplement (Supplemental Table S4).

Time to first relapse and time to CDP

Average treatment effect. The RMTL to first relapse in a 5-year follow-up period was lower for DMF than TERI (1.29 and 1.58, respectively). The RMTL of DMF patients was significantly lower than that of TERI patients for time to first relapse by 18.3% ($p = 0.001$) before adjustment and 25.7% ($p = 0.018$) after doubly robust adjustment of selected baseline variables (Table 2). The RMTL to CDP in a 5.1-year follow-up period was 0.921 for DMF and 0.996 for TERI. The RMTL of DMF patients was 7.5% and 5.5% lower than that of TERI patients for time to CDP before and after adjustment, respectively, with no statistical significance ($p > 0.05$, Table 2). All results suggested later onset of first relapse and CDP among DMF patients compared with TERI patients.

Performance of the ITR score and treatment effect heterogeneity. For time to first relapse, there was moderate treatment effect heterogeneity as reflected by the positive slopes of all validation curves in Figure 1. In subgroups containing a smaller proportion of patients with the lowest ITR scores (left side of the X-axis), the ATEs measured by both the RMTL ratio and HR ratio tended to be further below 1, implying stronger heterogeneity among high DMF responders. For time to CDP, the treatment effect heterogeneity was little as the validation curves in Figure 1 were relatively flat for both RMTL ratio and HR ratio. The validation curves agreed with the heterogeneity test for both time-to-event endpoints as shown in Table 2. Like ARR, boosting method identified smaller treatment effect heterogeneity compared with the other three scoring methods. See more information related to the two time-to-event endpoints in the Supplement.

The resulting ITR scores for ARR and RMTL ratios were highly correlated (estimated correlation coefficient = 0.82), corroborating each other in identifying high DMF responders. This high correlation suggested that the optimal treatment that reduced the ARR may also likely delay the onset of first relapse of the same patients.

Discussion

Previous related studies often focused on RCTs,[1,11] did not have enough registry data for DMF and TERI,[12] and the prediction methods of individual treatment response remained to be traditional regression-based models.[12,13] To our knowledge, this is the first application of a rigorous approach for precision medicine generalized to real-world data between DMF and TERI, and one of the few attempts in MS to disentangle the prognostic role of baseline covariates from their possible

Table 4. Patient characteristics by subgroups defined by the ITR score estimated from contrast regression for ARR ratio DMF versus TERI.

Baseline characteristics	High DMF responders *n* (%) 1674 (60)	Equal responders *n* (%) 1117 (40)	SMD[a]	*p*-Value[b]
Female	1240 (74.1)	762 (68.2)	0.130	<0.001
Age, years	44.59 (10.0)	37.54 (10.6)	0.685	<0.001
Prior DMTs, No.	1.17 (0.94)	0.65 (0.90)	0.569	<0.001
Time since MS diagnosis, months	109.5 (88.9)	50.1 (62.6)	0.772	<0.001
Relapses in the previous 12 months, No.	0.35 (0.56)	0.59 (0.70)	0.379	<0.001
Relapses in the previous 24 months, No.	0.57 (0.80)	0.86 (0.95)	0.332	<0.001
New diagnosis	430 (25.7)	668 (59.8)	0.734	<0.001
Prior GA (yes/no)	1090 (65.1)	1059 (94.8)	0.799	<0.001
Prior interferon (yes/no)	701 (41.9)	686 (61.4)	0.398	<0.001
EQ5D5L VAS score	86.7 (12.5)	84.3 (14.0)	0.177	<0.001
EDSS total score	1.62 (1.41)	2.35 (1.54)	0.490	<0.001
EDSS visual score	0.22 (0.53)	0.38 (0.76)	0.239	<0.001
EDSS ambulatory score	0.16 (0.80)	0.36 (1.17)	0.190	<0.001
EDSS brainstem score	0.16 (0.43)	0.28 (0.59)	0.226	<0.001
EDSS cerebellar score	0.38 (0.74)	0.62 (0.94)	0.285	<0.001
EDSS cerebral score	0.29 (0.55)	0.76 (0.91)	0.636	<0.001
EDSS pyramidal score	0.73 (0.99)	0.97 (1.16)	0.223	<0.001
EDSS sensory score	0.67 (0.80)	0.97 (0.99)	0.333	<0.001
EDSS sphincteric score	0.23 (0.54)	0.33 (0.57)	0.177	<0.001

ARR: annualized relapse rate; DMF: dimethyl fumarate; DMT: disease-modifying therapy; EDSS: Expanded Disability Status Scale; EQ5D5L: EuroQol-5 Dimension 5-level version; GA: glatiramer acetate; ITR: individualized treatment response; SD: standard deviation; SMD: standardized mean or proportion difference; TERI: teriflunomide; VAS: visual analogue scale.
Data are reported as mean (SD) for continuous variables, and *n* (%) for categorical variables.
New diagnosis is defined as "yes" if one of the following conditions is satisfied: (1) having a time since diagnosis of <12 months or (2) having no prior DMTs; and "no" otherwise.
[a]Standardized mean or proportion difference (Cohen's *d* values): a value <0.2 is considered acceptable, between 0.2 and 0.5 considered as a moderate difference, between 0.5 and 0.8 as a significant difference, and >0.8 as a major difference.
[b]*p*-Values are from unpaired *t*-tests (or Wilcoxon rank sum test if non-normally distributed) for continuous variables and Pearson Chi-square (or exact test extensions in case of low frequency) for categorical variables.

role as treatment effect modifiers.[1,14] We found a significant benefit of DMF relative to TERI in the entire NTD patient sample and detected a significant responder subgroup, among which DMF was substantially better than TERI with respect to ARR and time to first relapse (but not time to CDP). The factors affecting the relative benefit of DMF versus TERI included patients' age, number of prior DMTs, month since diagnosis, prior GA, and EDSS cerebral score. Prior GA had the largest weight, and the positivity suggested that the ARR ratio (DMF versus TERI) was lower for patients without receiving GA. Among the remaining patients, the effects of DMF versus TERI were statistically inconclusive and a better powered study is needed to draw further conclusions on the significance of both subgroups.

It is crucial to keep the treatment comparator and study population in mind when comparing or generalizing precision medicine conclusions (Supplemental Figure S6). Age and MS duration might seem to have opposite effects on ARR between DMF and TERI compared with existing literature of RCTs such as the DEFINE/CONFIRM trials,[15,16] where DMF responders were younger and more recently diagnosed, whereas the high DMF responders were older and had a longer MS duration. However, they are not in conflict for two main reasons: (1) the comparators were different (NTD compared DMF versus TERI and DEFINE/CONFIRM compared DMF versus placebo; Supplemental Figure S5); and (2) NTD patients were different from the DEFINE/CONFIRM patients at baseline in terms of age, MS

duration, number of relapses in the prior year, prior treatment, and EDSS score. For example, NTD patients (mean age: 40 and 45 years, mean MS duration: 6 and 9 years, respectively, for DMF and TERI) tended to be relatively older with longer MS duration than DEFINE/CONFIRM patients (mean age < 40 years and mean MS duration < 6 years).

The study has several limitations. First, we assumed that all confounders were observed and included in the analysis, which may not be true. The presence of unmeasured confounders (e.g., MRI or cognition data for this study) can lead to poor ITR score and biased evaluation. Second, we considered ARR as a meaningful summary of a patient's response to treatment when patients with the same ARR can have different lengths of treatment. A shorter treatment could imply that the patient experienced negative treatment effects and quickly switched to an alternative treatment, but this was not necessarily reflected in the ARR ratio. Third, the study presented a set of analyses as a proof-of-concept example without necessarily optimizing or justifying all analytical choices. Different PS, outcome regression, imputation, ITR scoring methods, and baseline variables could lead to different results. Fourth, it is not clear why prior GA was identified as an important treatment effect modifier. The mechanistic relationship between prior GA and relapses can be complicated, and further research is needed. Last, internal validation was applied to avoid over-optimistic ITR scores, and external validation with a sufficient sample size was not possible due to data availability, but it would be needed to draw conclusions beyond the NTD cohort.

Conclusions

This study sets a solid basis to build on future research. Patients may switch from treatment to treatment according to their personal experience. A possible future research direction can be a dynamic treatment strategy to guide patients to select the optimal treatment at baseline and make subsequent adjustment according to their clinical history.

Acknowledgments

Cara Farrell, Excel Medical Affairs, copyedited and styled the manuscript per journal requirements. Biogen reviewed and provided feedback on the paper. The authors had full editorial control of the paper and provided their final approval of all content.

Author contributions

X Jiang interpreted data and drafted the manuscript. F Pellegrini and A Harrington designed the study, coordinated the analyses, interpreted the data, and revised the manuscript. L Tian, and M Zuercher developed the statistical algorithms, analyzed and interpreted data, and revised the manuscript. G Simoneau, W Castro-Borrero, P van Hoevell, C de Moor, A Bergmann, Y Heer, and S Braune interpreted data. All authors provided final approval of the manuscript for submission.

Data availability

The data used in this study are owned by the NeuroTransData Registry and sharing of the data is subject to their policies. Any reasonable requests for data access can be directed to the NeuroTransData Registry (www.neurotransdata.com).

Declaration of conflicting interests

X Jiang, G Simoneau, A Harrington, W Castro-Borrero, C de Moor, F Pellegrini are employees and former employees of and hold stock/stock options in Biogen. L Tian received consulting fees from Biogen. M Zuercher, P van Hoevell, and Y Heer are employees of Rewoso AG, Zürich, Switzerland. A Bergmann has received consulting fees from advisory board, speaker, and other activities for NeuroTransData; project management and clinical studies for and travel expenses from Novartis and Servier. S Braune has received honoraria from Kassenaerztliche Vereinigung Bayern and health maintenance organizations for patient care; honoraria for consulting, project management, clinical studies, and lectures and from Biogen, Eli Lilly, Merck, NeuroTransData, Novartis, Roche, and Thieme Verlag; honoraria and expense compensation as board member of NeuroTransData.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was sponsored by Biogen (Cambridge, MA, USA). Biogen funded editorial support in the development of this paper. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

ORCID iD

Wanda Castro-Borrero https://orcid.org/0009-0002-9018-4263

Supplemental material

Supplemental material for this article is available online.

Notes

a. A relapse was defined as new or recurrent neurologic symptoms lasting at least one-day accompanied by new objective neurologic findings.
b. Confirmed disability progression (CDP) events were defined as at least 0.5-point EDSS score increases for patients with baseline EDSS score greater than 5.5, and at least 1.0-point EDSS score increases for patients with baseline EDSS score 0–5.5.[7]

References

1. 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* 2020; 26: 1064–1073.
2. 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* 2021; 116: 335–352.
3. Zhao L, Tian L, Cai T, et al. Effectively selecting a target population for a future comparative study. *J Am Stat Assoc* 2013; 108: 527–539.
4. Simoneau G, Jiang X, Rollot F, et al. 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. *Mult Scler J Exp Transl Clin* 2022; 8: 20552173221116591.
5. NeuroTransData GmbH. [NTD company website]. https://www.neurotransdata.com/ (2022, accessed 8 September 2022).
6. Bergmann A, Stangel M, Weih M, et al. Development of registry data to create interactive doctor–patient platforms for personalized patient care, taking the example of the DESTINY system. *Front Digit Health* 2021; 3: 633427.
7. Braune S, Grimm S, van Hövell P, et al.; NTD Study Group. 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.
8. Wehrle K, Tozzi V, Braune S, et al. Implementation of a data control framework to ensure confidentiality, integrity, and availability of high-quality real-world data (RWD) in the NeuroTransData (NTD) registry. *JAMIA Open* 2022; 5: ooac017.
9. Holstiege J, Steffen A, Goffrier B, et al. Epidemiologie der multiplen sklerose-eine populationsbasierte deutschlandweite studie. https://www.versorgungsatlas.de/fileadmin/ziva_docs/86/VA-86-Multiple%20Sklerose-Bericht-V13_Cor..pdf (2017, accessed 5 September 2022).
10. Team R Core. *R: a language and environment for statistical computing*. http://www.R-project.org/ (2013, accessed 5 September 2022).
11. Chalkou K, Steyerberg E, Egger M, et al. A two-stage prediction model for heterogeneous effects of treatments. *Stat Med* 2021; 40: 4362–4375.
12. Kalincik T, Manouchehrinia A, Sobisek L, et al.; MSBase Study Group. Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. *Brain* 2017; 140: 2426–2443.
13. Stühler E, Braune S, Lionetto F, et al. Framework for personalized prediction of treatment response in relapsing remitting multiple sclerosis. *BMC Med Res Methodol* 2020; 20: 24.
14. Steyerberg EW and Claggett B. Towards personalized therapy for multiple sclerosis: limitations of observational data. *Brain* 2018; 141: e38.
15. Fox RJ, Miller DH, Phillips JT, et al.; CONFIRM Study Investigators. Placebo-controlled phase 3 study of oral BG-12 or glatiramer in multiple sclerosis. *N Engl J Med* 2012; 367: 1087–1097.
16. Gold R, Kappos L, Arnold DL, et al.; DEFINE Study Investigators. Placebo-controlled phase 3 study of oral BG-12 for relapsing multiple sclerosis. *N Engl J Med* 2012; 367: 1098–1107.