Impact of methodological choices in comparative effectiveness studies: application in natalizumab versus fingolimod comparison among patients with multiple sclerosis

M. Lefort1,2†, S. Sharmin3,4†, J. B. Andersen5†, S. Vukusic6,7,8, R. Casey6,7,8,9, M. Debouverie10, G. Edan11, J. Ciron12, A. Ruet13, J. De Sèze14, E. Maillart15, H. Zephir16, P. Labauge17, G. Defer18, C. Lebrun-Frenay19, T. Moreau20, E. Berger21, P. Clavelou22, J. Pelletier23, B. Stankoff24, O. Gout25, E. Thouvenot26, O. Heinzlef27, A. Al-Khedr28, B. Bourre29, O. Casez30, P. Cabre31, A. Montcuquet32, A. Wahab33, J. P. Camdessanche34, M. Maurousset35, H. Ben Nasr36, K. Hankiewicz37, C. Pottier38, N. Maubeuge39, D. Dimitri-Boulos40, C. Nifile41, D. A. Laplaud42,43, D. Horakova44, E. K. Havrdova44, R. Alroughani45, G. Izquierdo46, S. Eichau46, S. Ozakbas47, F. Patti48,49, M. Onofri50, A. Lugasès51,52, M. Terzi53, P. Grammond54, F. Grand'Maison55, B. Yamout56, A. Pra57,58, M. Girard57,58, P. Duquette57,58, C. Boz59, M. Trojano60, P. McCombe61,62, M. Slees63, J. Lechner-Scott64,65, R. Turkoglu66, P. Sola67, D. Ferraro67, F. Granella68,69, V. Shaygannejad70, J. Prevost71, D. Maimone72, O. Skibina73, K. Buzzard73, A. Van der Walt73, R. Karabudak74, B. Van Wijmeersch75, T. Csepany76, D. Spitaleri77, S. Vucic78, N. Koch-Henriksen79, F. Sellebjerg80, P. S. Soerensen80, C. C. Hilt Christensen81, P. V. Rasmussen82, M. B. Jensen83, J. L. Frederiksen84, S. Bramow80, H. K. Mathiesen85, K. I. Schreiber80, H. Butzkueven86,87,88, M. Magyari3,4,†, T. Kalincik3*† and E. Leray1,2*†

†M. Lefort, S. Sharmin and J.B. Andersen contributed equally to the manuscript as first author.
†M. Magyari, T. Kalincik and E. Leray contributed equally to the manuscript.
*Correspondence: tomas.kalincik@unimelb.edu.au; emmanuelle.leray@ehesp.fr
1 Arènes - UMR 6051, RSMS (Recherche sur les Services et Management en Santé) - U 1309, Univ Rennes, EHESP, CNRS, Inserm, Rennes, France
3 Department of Medicine, University of Melbourne, Melbourne, Australia
Full list of author information is available at the end of the article

© The Author(s) 2022. Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.
Abstract

Background: Natalizumab and fingolimod are used as high-efficacy treatments in relapsing–remitting multiple sclerosis. Several observational studies comparing these two drugs have shown variable results, using different methods to control treatment indication bias and manage censoring. The objective of this empirical study was to elucidate the impact of methods of causal inference on the results of comparative effectiveness studies.

Methods: Data from three observational multiple sclerosis registries (MSBase, the Danish MS Registry and French OFSEP registry) were combined. Four clinical outcomes were studied. Propensity scores were used to match or weigh the compared groups, allowing for estimating average treatment effect for treated or average treatment effect for the entire population. Analyses were conducted both in intention-to-treat and per-protocol frameworks. The impact of the positivity assumption was also assessed.

Results: Overall, 5,148 relapsing–remitting multiple sclerosis patients were included. In this well-powered sample, the 95% confidence intervals of the estimates overlapped widely. Propensity scores weighting and propensity scores matching procedures led to consistent results. Some differences were observed between average treatment effect for the entire population and average treatment effect for treated estimates. Intention-to-treat analyses were more conservative than per-protocol analyses. The most pronounced irregularities in outcomes and propensity scores were introduced by violation of the positivity assumption.

Conclusions: This applied study elucidates the influence of methodological decisions on the results of comparative effectiveness studies of treatments for multiple sclerosis. According to our results, there are no material differences between conclusions obtained with propensity scores matching or propensity scores weighting given that a study is sufficiently powered, models are correctly specified and positivity assumption is fulfilled.

Keywords: Effectiveness, Multiple sclerosis, Propensity score, Indication bias, Causal contrasts, Censoring, Positivity assumption

Background

Natalizumab [1, 2] and fingolimod [3, 4] are two high-efficacy treatments used in Relapsing Remitting Multiple Sclerosis (RRMS) patients. Interestingly, the comparative effectiveness studies comparing these therapies showed results that were somewhat inconsistent [5–9]. In particular, we focus on three studies which used data from three multiple sclerosis (MS) registries, with differences in methods and conclusions [5–7]. We have already shown that some of this variability can be attributed to differences between the study populations [10, 11]. In the present work, we focus on the impact of methodological choices on the results—in particular, the methods used to control treatment indication bias and to manage censoring in time-to-event analysis.

In the absence of randomized clinical trials, many decisions need to be made to conduct observational studies. In the framework of “target trial”, developed by Hernan and Robins, we will focus on two protocol components, first, the assignment procedure and, second, the causal contrast [12]. First, to emulate the random assignment, we need to adjust for all known confounders [12]. Propensity score (PS), utilized in several ways, is a popular instrument used to control indication bias effect on the results of comparisons of intervention [13, 14]. The studies in the Danish MS Registry and MSBase used PS matching [6, 7] while the study in OFSEP used PS weighting [5]. Second, attrition bias and informative censoring result from systematic differences in the follow-up duration between cohorts. Two causal contrasts, per-protocol and intention-to-treat, were considered to evaluate follow-up information. While the per-protocol framework includes only outcomes that were recorded while patients were exposed to the relevant intervention, intention-to-treat framework mitigates the risk of informed censoring, which is of particular importance where clinical outcomes between interventions are delayed [12, 15]. The per-protocol framework was originally used in the studies in the Danish MS Registry and MSBase [6, 7] while the intention-to-treat framework was used in the OFSEP study [5]. Moreover, the study in MSBase used pairwise censoring that consists of censoring data within each PS matched pair to the shorter of the recorded follow-up times within the pair, in order to balance the analysed follow-up time between the groups [16].

The objective of this empirical study is to elucidate the influence of methodological decisions on the results of a comparison of two potent interventions, using the example of natalizumab and fingolimod among patients with MS and combined data from three large clinical registries [5–7].
Methods

Data source
This study is a result of a collaborative project [11, 17]. Longitudinal demographic and clinical data were extracted from MSBase on 15th of May 2018 [18, 19]. The Danish MS Registry cohort included all patients treated with natalizumab or fingolimod from 1st of July, 2011 when fingolimod became available in Denmark, until 1st of March, 2018 [20, 21]. The OFSEP cohort included data from 27 French university hospitals extracted from the European Database for Multiple Sclerosis (EDMUS) software in July 2014 [22]. No patient from OFSEP was recorded in MSBase. Some Danish patients who were recorded both in MSBase and Danish MS Registry (2% of Danish MS Registry) have been excluded from MSBase and only considered in the Danish MS Registry.

Eligibility criteria
All patients were diagnosed with RRMS. The required disability follow-up consisted of: a recorded visit with Expanded Disability Status Scale (EDSS)[23] score assessment within six months before treatment initiation (the baseline visit), two post-baseline visits with EDSS at least six months apart, and at least one on-treatment visit.

Interventions
Treatments of interest were the first exposure to natalizumab or fingolimod on or after 1st January 2011 and continued for a minimum of three months. Patients who participated in randomized trials or patients treated with off-label treatment (cyclophosphamide), or with therapies known to have extended duration of effect [24–26] (mitoxantrone, alemtuzumab, cladribine, daclizumab, rituximab, ocrelizumab) before the study therapy were excluded. Each patient could contribute only once to the follow-up analysis. When multiple eligible treatment starts were recorded, the earliest treatment was considered.

Outcomes
Four outcomes were evaluated to compare the relative effectiveness of the two study therapies:

1. Count of relapses.
2. Time to first relapse.
3. Time to first confirmed disability worsening event. Worsening was defined as an increase of ≥1.5 EDSS steps if baseline EDSS was 0, or 1.0 if baseline EDSS was 1.0–5.5, or 0.5 steps if baseline EDSS was >5.5, and sustained at all consecutive visits over ≥6 months (confirmation cannot be preceded by a relapse within 30 days).
4. Time to first confirmed disability improvement event. An improvement was defined as a decrease of 1.5 if baseline EDSS was 1.5, or 1.0 if baseline EDSS was 2.0–6.0, or 0.5 if baseline EDSS was >6, sustained at all consecutive visits over ≥6 months.

The end of analyzed study or period (count of relapses) depended on the definition of right-censoring (see below).

Assignment procedure: propensity score matching and weighting
In the present work, baseline was defined as the date of the start of the index therapy. To emulate the random assignment of treatments at baseline, PS [13, 27] was defined as the probability of being treated with natalizumab, conditional on the following baseline characteristics (based on expert opinion and prior analyses): sex, age, MS duration (from first MS symptoms to baseline), EDSS score, number of previous treatments, and, evaluated in the past 12 months: number of relapses, and, the nature of clinical activity recorded (disability worsening only, relapses only, both or no clinical activity). Country was added as random effect. We estimated both the average treatment effect for the treated (ATT) which is the average treatment effect among those patients who were exposed to natalizumab, and the average treatment effect for the entire eligible population (ATE) [28]. One-to-one, greedy, nearest neighbor, random matching on PS was used, allowing for approximating ATT only [29]. Matching caliper values of 0.1 (used in the original studies), 0.2 (as recommended by literature [30]) and 0.02 standard deviations of the PS (to prioritize close matching) were used. Two weighting procedures were explored. First, using Inverse Probability of Treatment Weighting (IPTW), the weights for a treated patient and for a control are defined as $w_i = \frac{1}{p_i}$ and $w_i = \frac{1}{1-p_i}$, respectively, where $p_i$ is the PS for a patient $i$. In order to reduce issue due to extreme weights, the weights were stabilized by multiplication by the marginal probability of receiving the treatment actually received [31], referred to as sIPTW. Second, using odds [32], the weight for a treated patient is 1 and the weight for control is defined $w_i = \frac{p_i}{1-p_i}$. Weighting with IPTW allows estimation of ATE while weighting by the odds allows estimation of ATT.

Causal contrast of interest
Intention-to-treat analysis retained all matched or weighted patients in the group as initial treatment allocation regardless of their following exposure, until either the last data entry or the study outcome. Per-protocol analysis retained all matched or weighted patients until the date of treatment discontinuation (or the date of last data entry if it occurs earlier). Pairwise-censoring was used as a technique of censoring after matching. In each pair, study follow-up of both patients was censored when
the follow-up of one of the two patients was censored. This approach prevented imbalance due to differential duration of follow-up in the matched groups.

**Sensitivity analysis without the positivity assumption**

The primary analysis ensured that the positivity assumption was fulfilled by only including patients who commenced natalizumab or fingolimod after the more recent of the two therapies became available on 1st January 2011. In a sensitivity analysis, all patients who commenced a study therapy were included, irrespective of the commencement date. Therefore, patients that were considered as ineligible in the primary analysis were included in this sensitivity analysis. Before 2011, MS patients had no chance to receive fingolimod, and could only started natalizumab; that is why the positivity assumption was violated.

**Statistical analysis**

Characteristics of the patients included in the analyses as well as those excluded by the matching procedure were described – overall and by treatment groups, before and after PS matching/weighting. Standardized mean differences (SMD) or Mahalanobis distances were computed, with 10% considered to be an acceptable difference [33]. Incidence of relapses was evaluated using a negative binomial model, with an offset term for follow-up durations. The cumulative hazards of first relapse, first EDSS improvement and first EDSS worsening were studied using Cox proportional hazards models with robust estimation of variance [34]. The models were either weighted by sIPTW or odds, or matched on PS. A cluster term (generalized estimating equations with negative binomial distribution) or a frailty term (Cox models) for pair identifier was used. As the probability of disability worsening and improvement events is associated with the frequency of EDSS scores [35], models with time to disability outcomes were adjusted for annualized visit density. All analyses were conducted for both the intention-to-treat and the per-protocol causal contrasts. Analyses using matching were completed with and without pairwise-censoring. Table 1 gives an overview of all the analytical approaches considered in the present work. The analyses were performed using R-software (R 3.4.0).

**Results**

**Patients’ characteristics**

Overall, 5,148 patients were included in this study [10]; 1,989 (39%) were treated with natalizumab and 3,159 (61%) with fingolimod. Patient’s characteristics are described in Table 2 (overall median age at baseline: 37.7 years; median MS duration at baseline: 6.9 years). Most of the patients had a clinically active disease and 70% had a baseline EDSS score equal or greater than 2. Table 3 presents the median durations of follow-up (overall: 3.1 years (interquartile range (IQR): 2.0–4.5)). The median durations of natalizumab and fingolimod treatments were 2.00 (1.3–3.1) and 2.2 (1.2–3.6) years, respectively.

**Patients’ characteristics after propensity score balancing procedures (matching and weighting)**

The distributions of PS showed a good overlap between the treatment groups, except in the tails (Fig. 1). The use of three caliper values for PS-matching led to three similar matched datasets (Table 2). The characteristics of the matched groups were comparable to the characteristics of the overall sample. The excluded patients tended to experience less disease activity. Table 4 presents patients’ characteristics by treatment group. Overall, 35% of patients treated with fingolimod had an EDSS score < 2 at treatment start while it was 22% in the group treated with natalizumab. The matching procedure improved the

| Outcome                              | PS method                  | Model                                                                 |
|--------------------------------------|----------------------------|----------------------------------------------------------------------|
| **Counts of relapses**               | Weighting                  | Weighted negative binomial model of disease outcomes by treatment     |
|                                      | Matching                   | Generalized estimating equations with negative binomial distribution and cluster for treatment status |
| **Time to first relapse**            | Weighting                  | Weighted Cox model of disease outcomes by treatment                   |
|                                      | Matching                   | Frailty Cox model of disease outcomes by treatment                    |
| **Time to first EDSS worsening**     | Weighting                  | Weighted Cox model of disease outcomes by treatment adjusted for visit density |
|                                      | Matching                   | Frailty Cox model of disease outcomes by treatment adjusted for visit density |
| **Time to first EDSS improvement**   | Weighting                  | Weighted Cox model of disease outcomes by treatment adjusted for visit density |
|                                      | Matching                   | Frailty Cox model of disease outcomes by treatment adjusted for visit density |

*a Analyses were conducted in intention-to-treat, on treatment and pairwise-censoring (matching only) frameworks

*b Two type of weights were considered (inverse probability weighting and weighting by the odds)

*c Three values of calipers were considered (0.02, 0.1, 0.2)
Table 2  Baseline characteristics of the overall study population, as well as the subgroups of patients unmatched and matched within different calipers

| All | Matching with caliper = 0.1 | Matching with caliper = 0.2 | Matching with caliper = 0.02 |
|-----|---------------------------|---------------------------|---------------------------|
| N = 5148 | N = 3258 | N = 1890 | N = 3278 | N = 1870 | N = 3232 | N = 1916 |
| **Sex** | | | | | | |
| Female | 3698 (72%) | 2342 (72%) | 1356 (72%) | 2352 (72%) | 1346 (72%) | 2332 (72%) | 1366 (71%) |
| Male | 1450 (28%) | 916 (28%) | 534 (28%) | 926 (28%) | 524 (28%) | 900 (28%) | 550 (29%) |
| **Age at treatment start** | 37.7 (30.1–44.7) | 37.3 (30.1–44.3) | 38.5 (31.8–45.6) | 37.2 (30.1–44.4) | 38.7 (31.8–45.4) | 37.4 (30.2–44.5) | 38.3 (31.5–45.1) |
| **MS duration at treatment start** | 6.9 (3.1–12.5) | 6.4 (2.7–2.0) | 7.9 (4.0–13.4) | 6.4 (2.6–2.0) | 7.9 (4.0–13.4) | 6.3 (2.6–12.0) | 8.1 (4.0–13.3) |
| **EDSS at treatment start** | | | | | | |
| Less than 2 | 1556 (30%) | 782 (24%) | 774 (41%) | 810 (25%) | 746 (40%) | 789 (24%) | 767 (40%) |
| Between 2 and 3.5 | 2384 (46%) | 1609 (49%) | 775 (41%) | 1593 (49%) | 791 (42%) | 1588 (49%) | 796 (42%) |
| 4 or more | 1208 (23%) | 867 (27%) | 341 (18%) | 875 (27%) | 33 (18%) | 855 (26%) | 353 (18%) |
| **Number of relapses in the previous 12 months** | | | | | | |
| 0 | 1857 (36%) | 1063 (33%) | 794 (42%) | 1085 (33%) | 772 (41%) | 1059 (33%) | 798 (42%) |
| 1 | 2021 (39%) | 1290 (40%) | 731 (39%) | 1268 (39%) | 753 (40%) | 1276 (39%) | 745 (39%) |
| 2 | 975 (19%) | 690 (21%) | 285 (15%) | 707 (22%) | 268 (14%) | 696 (22%) | 279 (15%) |
| 3 or more | 295 (6%) | 215 (7%) | 80 (4%) | 218 (7%) | 77 (4%) | 107 (7%) | 94 (5%) |
| **Number of previous MS treatments** | | | | | | |
| 0 | 836 (16%) | 582 (18%) | 254 (13%) | 580 (18%) | 256 (14%) | 584 (18%) | 252 (13%) |
| 1 | 2559 (50%) | 1594 (49%) | 965 (51%) | 1597 (49%) | 962 (51%) | 1558 (48%) | 1001 (52%) |
| 2 | 1187 (23%) | 733 (23%) | 449 (24%) | 744 (23%) | 443 (24%) | 748 (23%) | 439 (23%) |
| 3 or more | 566 (11%) | 344 (11%) | 222 (12%) | 357 (11%) | 209 (11%) | 342 (11%) | 224 (12%) |
| **MS activity in the previous 12 months** | | | | | | |
| None | 1438 (28%) | 776 (24%) | 662 (35%) | 782 (24%) | 656 (35%) | 764 (24%) | 674 (35%) |
| Worsening | 419 (8%) | 287 (9%) | 132 (7%) | 303 (9%) | 116 (6%) | 295 (9%) | 124 (6%) |
| Relapse | 2159 (42%) | 1395 (43%) | 764 (40%) | 1398 (43%) | 761 (41%) | 1397 (43%) | 762 (40%) |
| Relapse and worsening | 1132 (22%) | 800 (25%) | 332 (18%) | 795 (24%) | 337 (18%) | 776 (24%) | 356 (19%) |
| **Data source** | | | | | | |
| MS Base | 3293 (64%) | 1874 (58%) | 1419 (75%) | 1882 (57%) | 1411 (75%) | 1852 (57%) | 1441 (75%) |
| DMSR | 1444 (28%) | 1167 (36%) | 277 (15%) | 1179 (36%) | 265 (14%) | 1153 (36%) | 291 (13%) |
| OFSEP | 411 (8%) | 217 (7%) | 194 (10%) | 217 (7%) | 194 (10%) | 227 (7%) | 184 (10%) |

* N (%)  
** Median (Quartiles)

Table 3  Follow-up duration according to the outcomes of interest (in years)

| Outcome | Intention-to-treat analysis | Per-protocol analysis |
|---------|-----------------------------|----------------------|
| Counts of relapses | 3.17 [2.01; 4.59] | 2.09 [1.24; 3.41] |
| Time to first relapse | 3.11 [3.05; 3.18] | 2.27 [2.21; 2.31] |
| Time to first EDSS worsening | 3.16 [3.10; 3.23] | 2.11 [2.07; 2.16] |
| Time to first EDSS improvement | 3.20 [3.13; 3.27] | 2.08 [2.02; 2.12] |

* Median (Quartiles) length of follow-up  
** Median [95% confidence interval] survival time of the reverse Kaplan–Meier, taking into account the length and the completeness of follow-up
balance between the compared groups, except for the data source and the number of previous MS treatments.

Table 5 presents patients’ characteristics by treatment group after weighting on sIPTW or odds. The treatment groups were well balanced, with SMD or Mahalanobis distances around 10% for all patient characteristics, except for the number of previous MS treatments, as natalizumab tended to be prescribed as first treatment more frequently than fingolimod. Exposure following the study therapy is shown in Table S1.

Comparison of effectiveness between natalizumab and fingolimod

Figure 2 summarises the results of all comparative analyses. While the estimated 95% confidence intervals of the estimated differences between natalizumab and fingolimod largely overlapped in all analyses, some variation in point estimates was observed.

With a few exceptions, the results of the analyses with matching and weighting led to the same conclusions, i.e., superiority of natalizumab (for relapse outcomes and EDSS improvement) or no evidence of difference (for EDSS worsening). Inconsistencies were observed mainly in the intention-to-treat frameworks, for relapse counts and first EDSS improvement. Weighting by the odds (ATT) tended to provide lower point estimates and similar margins of error of the relative effect compared to weighting by sIPTW (ATE). The value of the matching caliper did not influence the magnitude of the estimated differences.

Most of the variability in the estimates was linked to the causal contrast. The intention-to-treat paradigm led to less stable results, especially for the count of relapses and first EDSS improvement. For all outcomes except time to first EDSS worsening, the intention-to-treat analyses underestimated the differences between the therapies in comparison to per-protocol analyses with or without pairwise-censoring. Per-protocol analyses and pairwise-censored analyses returned similar point estimates, even though the margin of error varied. In the pairwise-censored analyses, confidence intervals were relatively smaller for relapse counts but larger for the disability outcomes compared to the per-protocol analysis.

Sensitivity analysis: positivity assumption

To test the effect of violation of the positivity assumption, 7,118 patients were included irrespectively of the date of their treatment start, of whom 3,726 were treated with natalizumab. The other baseline characteristics were similar to those of the main cohort (Table S3). The PS distribution was left-skewed in patients who commenced natalizumab before fingolimod became available (Figure S1). Using weighting, the comparison of the treatment effects on relapses was similar to the main analysis (Table 6). However, the point estimates for the difference in the treatment effects on EDSS worsening were substantially lower than in the primary analysis, although confidence intervals overlapped. When matching was used, the estimates for EDSS outcomes were less influenced by the violation of the positivity assumption. Nevertheless, the estimates of the differences between treatment effects on relapses were substantially inflated when the assumption was violated, especially for the intention-to-treat causal effect.

Discussion

In this empirical study conducted on a complex chronic neurological condition, with long-term follow-up data, several non-linear outcomes and well powered dataset, most of the methodological choices (PS matching/weighting, caliper values, weighting on IPTW vs. odds, and pairwise censoring) resulted in consistent overall conclusions, in accordance with two of the three original studies [5, 6], the pooled analysis [11] and a recent French head-to-head prospective study [36]. In a longitudinal observational study conducted over the long-term in the presence of frequent changes of therapy, an intention-to-treat causal contrast tends to be associated with more variability in the observed effects than a per-protocol contrast. Importantly, violation of the positivity assumption demonstrated the most pronounced negative effect on the consistency of reported results.
Table 4  Characteristics at baseline according to treatment group in the overall population and when three matching calipers were used

|                | Overall | Matching (caliper = 0.1) | Matching (caliper = 0.2) | Matching (caliper = 0.02) |
|----------------|---------|--------------------------|--------------------------|---------------------------|
|                | N = 5148 | N = 3258                 | N = 3278                 | N = 3232                  |
|                | natalizumab | fingolimod | SMDc | natalizumab | fingolimod | SMDc | natalizumab | fingolimod | SMDc | natalizumab | fingolimod | SMDc |
|                | N = 1989 | N = 3159 |       | N = 1629 | N = 1629 |       | N = 1639 | N = 1639 |       | N = 1616 | N = 1616 |       |
| Sexa           |         |           |       |         |           |       |         |           |       |         |           |       |
| Female         | 1451 (73%) | 2247 (71%) |       | 1175 (72%) | 1167 (72%) |       | 1183 (72%) | 1169 (71%) |       | 1169 (72%) | 1163 (72%) |       |
| Male           | 538 (27%) | 912 (29%) |       | 454 (28%) | 462 (28%) |       | 456 (28%) | 470 (29%) |       | 447 (28%) | 453 (28%) |       |
| Age at treatment startb | 36.6 (29.3; 43.9) | 38.5 (31.6; 45.4) | 13% | 37.2 (29.8–44.4) | 37.4 (30.3–44.2) | 0.5% | 37.2 (29.7–44.4) | 37.4 (30.3–44.3) | 1% | 37.3 (29.7–44.4) | 37.6 (30.5–44.5) | 3% |
| MS duration at treatmentb | 6.3 (2.4; 11.8) | 7.4 (3.6; 13.0) | 13% | 6.3 (2.3–12.2) | 6.5 (2.9–11.8) | 2% | 6.2 (2.3–12.2) | 6.6 (2.9–11.9) | 2% | 6.3 (2.3–12.2) | 6.4 (2.8–11.9) | 2% |
| EDSS at treatment startb |          |           |   |         |           |       |         |           |       |         |           |       |
| 2 or less      | 434 (22%) | 1122 (33%) |       | 374 (23%) | 408 (25%) |       | 377 (23%) | 433 (26%) |       | 372 (23%) | 417 (26%) |       |
| Between 2 and 4| 981 (49%) | 1403 (44%) |       | 822 (50%) | 787 (48%) |       | 836 (50%) | 767 (47%) |       | 813 (50%) | 775 (48%) |       |
| 4 or more      | 574 (29%) | 634 (20%) |       | 433 (27%) | 433 (27%) |       | 436 (25%) | 439 (27%) |       | 431 (27%) | 424 (26%) |       |
| Number of relapses in the previous 12 monthsa |          |           |   |         |           |       |         |           |       |         |           |       |
| 0              | 570 (29%) | 1287 (41%) |       | 543 (33%) | 520 (32%) |       | 545 (33%) | 540 (33%) |       | 541 (33%) | 518 (32%) |       |
| 1              | 752 (38%) | 1269 (40%) |       | 620 (38%) | 670 (41%) |       | 623 (38%) | 645 (39%) |       | 618 (38%) | 658 (41%) |       |
| 2              | 464 (24%) | 491 (15%) |       | 346 (21%) | 344 (21%) |       | 351 (22%) | 356 (22%) |       | 350 (22%) | 346 (21%) |       |
| 3 or more      | 183 (9%) | 112 (9%) |       | 120 (7%) | 95 (6%) |       | 120 (7%) | 98 (6%) |       | 107 (7%) | 94 (6%) |       |
| Number of previous MS treatmentsa |          |           |   |         |           |       |         |           |       |         |           |       |
| 0              | 401 (20%) | 435 (14%) |       | 334 (21%) | 248 (15%) |       | 337 (21%) | 243 (15%) |       | 334 (21%) | 250 (15%) |       |
| 1              | 924 (46%) | 1635 (52%) |       | 742 (46%) | 852 (52%) |       | 746 (46%) | 851 (52%) |       | 732 (44) | 826 (51%) |       |
| 2              | 457 (23%) | 730 (23%) |       | 367 (23%) | 371 (23%) |       | 370 (23%) | 374 (23%) |       | 365 (23%) | 383 (24%) |       |
| 3 or more      | 207 (10%) | 359 (11%) |       | 186 (11%) | 158 (10%) |       | 186 (11%) | 171 (10%) |       | 185 (11%) | 157 (10%) |       |
| MS activity in the previous 12 monthsa |          |           |   |         |           |       |         |           |       |         |           |       |
| None           | 410 (21%) | 1028 (32%) |       | 393 (24%) | 383 (24%) |       | 395 (24%) | 387 (24%) |       | 390 (24%) | 374 (23%) |       |
| Worsening      | 160 (8%) | 259 (8%) |       | 150 (9%) | 137 (8%) |       | 150 (9%) | 153 (9%) |       | 151 (9%) | 144 (9%) |       |
| Relapse        | 886 (44%) | 1273 (40%) |       | 686 (42%) | 709 (44%) |       | 694 (42%) | 704 (43%) |       | 690 (43%) | 707 (44%) |       |
| Relapse and worsening | 533 (27%) | 599 (19%) |       | 400 (25%) | 400 (25%) |       | 400 (24%) | 395 (24%) |       | 385 (24%) | 391 (24%) |       |
| Data sourcea  |          |           |   |         |           |       |         |           |       |         |           |       |
| MS Base        | 1141 (57%) | 2152 (68%) |       | 949 (58%) | 925 (57%) |       | 957 (58%) | 925 (56%) |       | 935 (58%) | 917 (57%) |       |
| DMSR           | 607 (31%) | 837 (26%) |       | 607 (37%) | 560 (34%) |       | 607 (37%) | 572 (35%) |       | 593 (37%) | 560 (35%) |       |
| OFSEP         | 241 (12%) | 170 (5%) |       | 73 (4%) | 144 (9%) |       | 75 (5%) | 142 (9%) |       | 88 (5%) | 139 (9%) |       |

*a N (%)  
*b Median (Quartiles)  
*c SMD standardized mean differences or Mahalanobis distances between Natalizumab treated patients and Fingolimod treated patients
Propensity score to control indication bias

Among the four methods using PS, matching and weighting have shown a superior performance to adjustment and stratification in achieving balance on baseline characteristics [37], reduction of bias and estimation of variance [38–40]. Therefore, we restricted our present work to PS matching and weighting. The results of the weighting and matching procedures were consistent, confirming that both methods performed well in sufficiently powered data sets and correctly specified models. The width of the matching caliper did not have much influence on the consistency of the results, confirming that 0.2 is a sufficiently conservative caliper, as previously reported [30]. The only detectable systematic variability

Table 5 Characteristics at baseline by treatment group in the overall study sample, and cohorts weighted on sIPTW and odds

|                | Unweighted | Weighting using sIPTW | Weighting using the odds |
|----------------|------------|-----------------------|-------------------------|
|                | natalizumab | fingolimod | SMD<sup>c</sup> | natalizumab | fingolimod | SMD<sup>c</sup> | natalizumab | fingolimod | SMD<sup>c</sup> |
|                | N = 5148   | N = 1989   | N = 3159 | N = 3159   | N = 1989   | N = 3159   | N = 3159   | N = 3159   |
| Sex<sup>a</sup> |            |            |            |            |            |            |            |            |            |
| Female         | 1451 (73%) | 2247 (71%) | 71%       | 73%       | 72%       | 73%       | 73%       | 73%       |
| Male           | 538 (27%)  | 912 (29%)  | 29%       | 27%       | 28%       | 27%       | 27%       | 27%       |
| Age at treatment start<sup>b</sup> | 36.6 (29.3; 43.9) | 38.5 (31.6; 45.4) | 13% | 37.6 (30.4–45.2) | 37.9 (30.7–44.8) | 2% | 36.6 (29.3–43.9) | 36.8 (29.8–43.9) | 1% |
| MS duration at treatment<sup>b</sup> | 6.3 (2.4; 11.8) | 7.4 (3.6; 13.0) | 13% | 6.8 (2.7–12.9) | 7.0 (3.2–12.4) | 2% | 6.2 (2.4–11.8) | 6.2 (2.7–11.6) | 1% |
| EDSS at treatment start<sup>c</sup> | 32% | 12% | 4% | 23% | 23% | 23% | 23% | 23% |
| Number of relapses in the previous 12 months<sup>a</sup> | 37% | 4% | 3% | 37% | 4% | 3% | 37% | 4% | 3% |
| Number of previous MS treatments<sup>a</sup> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Number of previous MS treatments<sup>a</sup> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| MS activity in the previous 12 months<sup>a</sup> | None | Worsening | Relapse | Relapse and worsening | None | Worsening | Relapse | Relapse and worsening | None | Worsening | Relapse | Relapse and worsening |
| Data source<sup>a</sup> | MS Base | DMSR | OFSEP |
| N (%) | 57 (57%) | 2152 (68%) | 57% | 53% |
| Median (Quartiles) | 259 (8%) | 42% | 44% | 43% |
| SMD<sup>c</sup> | 2% | 2% | 2% | 2% | 2% | 2% | 2% | 2% | 2% | 2% | 2% | 2% |
| SMD<sup>c</sup> | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% | 3% |

<sup>a</sup> N (%)  
<sup>b</sup> Median (Quartiles)  
<sup>c</sup> SMD standardized mean differences and Mahalanobis distance between natalizumab treated patients and fingolimod treated patients
was noted for the type of estimated effect, with the magnitude of the ATE effect trending towards higher values for relapse incidence and time to first relapse.

The matched study sample corresponds to an overlap between the fingolimod- and the natalizumab-treated target populations, with inclusion of comparable cases and exclusion of cases outside the common distribution of the PS (ATT effect of interest). Such reductions in sample size may lead one to study a very specific subpopulation and, so, impact the precision and the generalizability of the results [41]. An IPTW-weighted sample is closer to the entire study population, especially where ATE is the effect of interest. It is therefore not surprising, given that the use of natalizumab and fingolimod...
in MS differs in clinical settings, that we have observed differences in the point estimates obtained with the matched and weighted analyses. Weighting could potentially be subject to influential cases with extreme weights, which are excluded from matching, as they fall outside of the central portion of the PS distribution [42]. In this work, we used stabilized weights to mitigate the risk of influential cases, as an alternative to weight trimming or truncation [33].

Management of censoring

In the present study, most irregularities were related to the intention-to-treat causal contrast, which resulted in less stable and often deflated estimates than the per-protocol analysis. These fluctuations were more pronounced for the outcomes defined as counts of events and time to medium-term events (first disability worsening or improvement) than for time to short-term events (first relapse). The intention-to-treat exchanges the association with the outcome, irrespective of treatment status over-time, and addresses the question of the effect of treatment decision, irrespective of further persistence on the assigned therapy. Therefore, such an approach leads to conservative estimates, which explains the observed overall deflation of effect sizes in comparison to the per-protocol approach and the minimum impact on short-term outcomes.

On the other hand, patients and neurologists may be more interested in a per-protocol effect, which estimates the effect of an intervention while being adhered to. However, a per-protocol treatment effect can be inflated by attrition bias and informed censoring, especially when one of the compared interventions is a-priori perceived as being more effective [43]. This would lead to the selection of “treatment responders”, because patients who respond well to treatment are more likely to remain treated than non-responders [44]. In addition, the per-protocol requirement of adherence to treatment may introduce additional selection bias, which may limit generalizability of conclusions [45], whereas the intention-to-treat approach preserves the balance established at baseline. A pairwise-censoring procedure can be combined with either causal contrast. Its purpose is to sustain the balance between the matched cohorts even when censoring / treatment cessation is systematically different between the compared groups. This sustained balance is achieved at the expense of loss of part of study follow-up due to right-censoring of the paired cases. However, in the present empirical analysis, per-protocol and pairwise-censored analyses led to similar conclusions and point estimates. The observed increase in the margin of error in pairwise-censored analysis suggests some loss of power. Marginal structural models with IPTWs accounting for the probability of censoring may provide a more efficient solution, as they do not lead to loss of follow-up information [46–48].

Positivity assumption

The positivity assumption can be objectively assessed in several steps. First, the definition of study timeline and area should be such as both treatments are available to all included patients. Second, the common support of PS distribution in the two groups needs to be established [31]. In our main analysis, these two steps confirmed that the positivity assumption was met. To examine the importance of the positivity assumption, in a different analysis, we allowed inclusion of patients before one of the studied therapies (fingolimod) became available. This included more natalizumab-treated patients from a time period when the probability of exposure to fingolimod was zero. The results of this analysis showed the most pronounced variability and the largest deviation from the primary analysis. Therefore, in a sufficiently powered longitudinal dataset, non-zero probability of exposure to both compared therapies at all baseline time-points is the most important aspect of methodological considerations explored in this study.

Limitations

Through consistency and exchangeability assumptions, it is assumed that there were no unmeasured confounders. Nevertheless, our study was limited by incomplete MRI data, while MRI activity is a known prognostic factor in MS [49]. Reassuringly, two of our three previous studies that accounted for MRI at treatment start showed results consistent with our primary analysis [5, 6].

In addition, heterogeneity of data in multisite registries (with potential differences in therapeutic practices, health care systems and treatment access) may increase variance of the associations between treatments and outcomes [50]. On the other hand, heterogeneity that is representative of clinical use of the compared therapies extends generalizability of the results. We have mitigated the potential heterogeneity in the present dataset by including country as a random term in the PS modeling. Finally, this study did not attempt to compare the efficiency and robustness of different analytical methods, as this can be done only with simulation studies. Instead, we have focused on the evaluation of practical methodological questions in the context of a specific clinical choice.

Conclusion

This empirical study provides practical insights into the effects of several methodological choices on the estimates of the difference between two therapies in the context of a chronic neurological disease, in a sufficiently powered analysis and correctly specified models. Our results lead us to conclude that methodological considerations such as PS matching/weighting and their specifications, causal contrast and
management of censoring have a negligible effect on the overall analyses, given that the model assumptions are met. The choice between ATT or ATE as the preferred approach should be driven by the clinical question of interest. In our clinical example, when both treatments can be prescribed to patients with relapsing–remitting MS following similar rules, there is no apparent reason to restrict the analysis to the natalizumab- or the fingolimod-treated patients, and ATE may be the preferred estimator of interest.

A recent review highlighted the good practice in the use and reporting of PS in MS [41]. While methodological choices in observational studies remain challenging, our present work illustrates the priorities for methodological aspects of PS-based analyses of comparative treatment effectiveness in large registries.

Abbreviations
ATT: Average Treatment effect for the Treated; ATE: Average Treatment effect for the Entire eligible population; EDMUS: European Database for Multiple Sclerosis; ESSS: Expanded Disability Status Scale; IPTW: Inverse Probability of Treatment Weighting; MS: Multiple Sclerosis; PS: Propensity Score; RRM: Relapsing Remitting Multiple Sclerosis; sIPTW: Stabilized Inverse Probability of Treatment Weighting; SMD: Standardized Mean Differences.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01623-8.

Additional file 1: Figure S1. Propensity score distribution without the positivity assumption. Table S1. Treatment exposure after natalizumab or fingolimod start during the follow-up. Table S2. Baseline characteristics of the unmatched cohorts by treatment group. Table S3. Baseline characteristics of cohort violating the positivity assumption.

Acknowledgements
Not applicable

Authors’ contributions
ML, SS, and JBA contributed to the design of the study, conducted and interpreted the analysis, and drafted, revised, and approved the manuscript. HB, MM, TK, and EL conceptualized and designed the study, contributed to data acquisition, interpreted the results, and revised and approved the manuscript. SV, NKH, RC, and DAL contributed to the design of the study, interpreted the results, and revised and approved the manuscript. MD, GE, JC, AR, JDS, EM, HZ, PL, GD, CLF, TM, EB, PC, JP, BS, OG, ET, OH, AA, BB, OC, PC, AM, AW, JPC, AM, HBN, KH, CP, NM, DDB, CN, DH, EKH, RA, GI, SE, SO, FP, MO, AL, MT, PG, FG, BY, AR, MG, PD, CB, MT, PMC, MS, JLS, RT, PS, DF, FG, VS, JP, DM, OS, KB, AW, RK, BVW, TC, DS, SV, FS, PSS, CCHC, PVR, MBJ, JLF, SB, HKM, and KS, contributed substantially to data acquisition and interpretation of the analysis, and revised and approved the manuscript. ML, SS, and JBA contributed equally. HB, MM, TK, and EL contributed equally. The authors read and approved the final manuscript.

Funding
OFSEP was 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, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation. ML has received travel grant from ARSEP foundation for this project. The Clinical Outcomes Research unit at the University of Melbourne received funding from NHMRC (grant number 1140766, 1129789, and 1157717) to support this study. The MSBase Foundation is a not-for-profit organization that receives support from Biogen, Novartis, Merck, Roche, Teva Pharmaceutical Industries and Sanofi Genzyme. The Danish Multiple Sclerosis Registry did not receive any funding to collaborate in this study.

Availability of data and materials
OFSEP: The individual data from the present study can be obtained upon request and after validation from the OFSEP scientific committee (see website: http://www.ofsep.org/en/dataaccess). MSBase: MSBase is a data processor, and warehouses data from individual principal investigators who agree to share their datasets on a project-by-project basis. Each principal investigator will need to be approached individually for permission to access the datasets. DMSR: Anonymized data will be shared on request from any qualified researcher under approval from the Danish Data Protection Agency.

Declarations
Ethics approval and consent to participate
OFSEP (Observatoire Français de la Sclérose en Plaques; French MS registry), ClinicalTrials.gov ID: NCT02889965, prospectively collects longitudinal data on clinical, biological, and imaging markers from patients who provided written informed consent following the French law on Bioethics. Storing data for research purposes was approved by the French Commission Nationale de l’Informatique et des Libertés (CNIL). MSBase is an international multiple sclerosis registry (World Health Organization International Clinical Trials Registry ID: ACTRN1260500455662) of observational data collected longitudinally as part of routine clinical care from 129 mostly tertiary multiple sclerosis centres in 36 countries. MSBase was approved by the Melbourne Health Human Research Ethics Committee and by the site institutional review boards, unless exemptions were granted according to the local regulations. Written informed consent was obtained from enrolled patients. The Danish Multiple Sclerosis Registry (DMSR) is a nationwide population-based registry consisting longitudinal data of all patients receiving disease-modifying treatments. Data are collected prospectively and stored following the data protection law of the Danish Data Inspection. The study obtained approvals from the Center for Data Review applications (j. nr. 2012–58-0004/V/2018–121-suit 6361). Consent has been obtained from each patient included in this study. Please see ‘Ethics statement’ in the manuscript for details. All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
OFSEP: The authors report the following relationships: speaker honoraria, advisory board or steering committee fees, independent data monitoring committees fee, and/or conference travel support from Actelion (PC, ET), Ad Scientiam (EM), Akcea (JPC), Atylaml (JPC), Almirall (OH), Bayer (GE, HZ, OH), Biogen (GE, JC, AR, JDS, EM, HZ, PL, GD, TM, EB, PC, JP, BS, ET, OH, BB, OC, AMo, JPC, AMa, IP, NM, DAL, SV), Celgene (JC, ET, DAL), CSL-Behring (JPC), FHI Imminent (HZ), Geneuro (SV), Genzyme-Sanoﬁ (GE, JC, AR, JDS, EM, HZ, PL, GD, CLF, TM, EB, PC, JP, BS, ET, OH, BB, OC, AMo, JPC, AMa, IP, NM, DAL, SV), GifROLS (JPC), Laboratoire Francais des Biotechnologies (JPC), LFBI (JPC), LFSEPS (HZ), Merck / EMDC (GE, JC, AR, EM, HZ, PL, GD, PC, JP, BS, ET, OH, BB, OC, AMo, JPC, AMa, NM, DAL, SV), Medday (EL, AR, TM, PC, JP, DAL, SV), Natus (JPC), Novartis (EL, GE, JC, AR, JDS, EM, HZ, PL, GD, CLF, TM, EB, PC, JP, BS, ET, OH, BB, OC, AMo, JPC, AMa, IP, NM, DAL, SV), Pfizer (JPC), Pharmarcia (JPC), Roche (ML, EL, GE, JC, AR, JDS, EM, HZ, PL, GD, CLF, EB, PC, JP, BS, ET, OH, BB, OC, AMo, JPC, AMa, NM, DAL, SV, WA), SNF-Floerger (JPC), Teva (GE, JC, AR, JDS, EM, HZ, PL, GD, EB, PC, JP, ET, OH, BB, AMo, JPC, AMa, SV), Académie de Médecine (HZ), Agence Nationale de la Recherche (DAL), French National Security Agency of Medicines and Health Products (EL), the EDMUS Foundation (EL), the ARSEP foundation (GE, HZ, ET, DAL, ML), PHRC Foundation (ET), Rennes University Hospital (GE). MSBase: The authors report the following relationships: speaker honoraria, advisory board or steering committee fees, research support and/or conference travel support from Actelion (EHK), Almirall (GI, FP, MT), Bayer (RA, FP, AL, MT, CB, MT, JLS, BVW, TC, DS), BioCsl (KB,
Multiple Sclerosis Over 15 Years. Neurology [Internet]. 2021 Feb 2;96(5):e783–e797. Available from: http://n.neurology.org/content/96/5/e783.abstract

49. Tintore M, Rovira À, Río J, Otero-Romero S, Arrambide G, Tur C, et al. Defining high, medium and low impact prognostic factors for developing multiple sclerosis. Brain. 2015;138(7):1863–74.

50. Bovis F, Signori A, Carmisciano L, Maietta I, Steinerman JR, Li T, et al. Expanded disability status scale progression assessment heterogeneity in multiple sclerosis according to geographical areas. Ann Neurol. 2018;84(4):621–5.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.