Ocrelizumab versus Natalizumab in Relapsing-Remitting Multiple Sclerosis: A Registry-Linked Electronic Health Records Study

Feiqing Huang, PhD[1,3], Wen Zhu, MD MS[6], Jue Hou, PhD[2], Sara Morini Sweet, MS[7], Yunqing Han, MS[7], Jun Wen, PhD[5], Katherine P. Liao, MD MPH[5,7], Tianrun Cai, MD[5,7], Tanuja Chitnis, MD[8], Florence T. Bourgeois, MD MPH[3,4], Zongqi Xia, MD PhD[6], Tianxi Cai, ScD [1,5]

1. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
2. Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, MN, USA.
3. Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA
4. Pediatric Therapeutics and Regulatory Science Initiative, Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA.
5. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
6. Department of Neurology, University of Pittsburgh, Pittsburgh, PA, USA.
7. Division of Rheumatology, Inflammation, and Immunity, Mass General Brigham, Boston, MA, USA.
8. Department of Neurology, Mass General Brigham, Boston, MA, USA.

Count:

Title (without subtitle):	13 words, 123 characters (with spaces)
Running Title:	18 characters (with spaces)
Abstract:	348 words
Manuscript:	3719 words
References:	42
Tables:	1
Figures:	3 (3 color figures)
Supplementary Tables:	7
Supplementary Figures:	3

Key Words: real-world evidence, electronic health records, imputation, confounding, multiple sclerosis, disability

Running Title: OCR vs NTZ in RRMS

Corresponding Authors:

Tianxi Cai, ScD

Department of Biomedical Informatics, Harvard Medical School

Address: 10 Shattuck Street, Room 434, Boston, MA 02115

Phone: 617-432-4923; Fax: 617-432-0693; Email: tcai@hsph.harvard.edu

Zongqi Xia, MD, PhD

Department of Neurology, University of Pittsburgh

Address: 3501 Fifth Ave, BST3-10.044

Phone: 412-383-5377; Fax: 412-648-7223; Email: zxia1@post.harvard.edu

ABSTRACT

Background: Ocrelizumab and natalizumab are commonly prescribed high-effectiveness disease-modifying therapies (DMTs) for relapsing-remitting multiple sclerosis (RRMS). However, no randomized clinical trial and few real-world studies have directly compared their effectiveness in reducing disability progression. Subtype classification and disability status are critical for multiple sclerosis (MS) research, but these data are often missing in electronic health records (EHRs), limiting robust real-world evidence generation.

Objective: To compare the effectiveness of ocrelizumab and natalizumab in two-year rater-assessed disability progression among RRMS patients using longitudinal registry-linked EHR data.

Design: Retrospective cohort study.

Setting: A large healthcare system that includes both academic and community practices.

Participants: Patients diagnosed with MS who initiated ocrelizumab or natalizumab between 2012 and 2020, with at least 6-month EHR data before treatment initiation and no prior exposure to other high-effectiveness DMTs.

Exposures: Treatment with ocrelizumab vs natalizumab.

Measurements: We developed an ensemble machine learning model to impute RRMS subtype and disability outcomes using structured and narrative EHR data. The primary outcome was moderate/severe rater-assessed disability at 2 years (observed or imputed Expanded Disability Status Scale [EDSS]≥4) after treatment initiation. We estimated the average treatment effects using semi-supervised doubly robust approach with comprehensive confounder adjustment and calibration to mitigate imputation bias. Covariates included standard demographic and clinical features such as baseline disability as well as knowledge graph-selected features. Sensitivity analyses used observed EDSS scores in registry-derived RRMS patients. Exploratory analyses

included rituximab, another B-cell–depleting therapy, with adjustments for differences in patient profiles.

Results: Among RRMS patients, those treated with ocrelizumab (n=543) had a significantly lower two-year risk of moderate/severe disability compared with those treated with natalizumab (n=205) based on imputed outcomes (risk difference, –5.87%; 95% CI: –11.28% to –0.46%; p=0.033) after confounder adjustment. Sensitivity analyses yielded consistent findings using imputed or observed EDSS outcomes in registry-derived RRMS patients.

Conclusion and relevance: In this real-world comparative effectiveness study using a novel semi-supervised doubly-robust framework, ocrelizumab was associated with a lower risk of disability progression than natalizumab among RRMS patients. This approach provides a roadmap for generating robust large-scale real-world evidence in settings of missing key inclusion features and outcomes.

INTRODUCTION

Multiple sclerosis (MS) is a chronic autoimmune disease that causes inflammatory demyelination in the central nervous system and progressive neurodegeneration, leading to disability.[1] Among more than 20 approved MS disease-modifying therapies (DMTs) with distinct mechanisms of action, robust comparative evidence for therapies of similar perceived effectiveness remains lacking. Within higher-effectiveness options, natalizumab [NTZ] and ocrelizumab [OCR] are widely prescribed for relapsing-remitting MS (RRMS) and associated with reduced relapse.[2–9] However, no randomized controlled trial (RCT) and few real-world studies have compared their long-term effects on disability.[5–10]

RRMS is more common than primary progressive MS (PPMS) and secondary progressive MS (SPMS) subtype.[11] MS subtype classification informs RCT eligibility, confounder adjustment in observational studies, and treatment decisions in clinical practice. While research registries often record MS subtype, it is rarely captured in electronic health records (EHR).[12]

The Expanded Disability Status Scale (EDSS) is a clinician-rated measure of neurological impairment across eight functional systems. Higher EDSS scores indicate greater disability.[13,14] Although EDSS is frequently used to evaluate disability outcome in RCTs, it has limitations. Like MS subtype, EDSS scores are typically documented in registries but not in EHRs.

Integrating registry and EHR data can strengthen real-world evidence by combining disease subtype and outcome measures (often not recorded during routine care) from registries with longitudinal clinical data besides the disease of interest (rarely recorded in disease registries) from EHR. Here, we present a methodological framework for linking registry and EHR data to phenotype MS subtype and EDSS outcomes beyond registry populations. As a clinically relevant

illustration, we compared the effectiveness of OCR versus NTZ in delaying two-year disability progression among patients with imputed or registry-derived RRMS subtype.

METHODS

Ethics Approval

The institutional review boards of the University of Pittsburgh (STUDY20070274) and Mass General Brigham (2023P001450) approved the study protocols. The use of de-identified clinical data was deemed exempt.

Data Sources and Patient Cohort

Figure 1 illustrated the overall study design. This study leveraged registry-linked EHR data.[15–19] EHR data from Mass General Brigham (MGB) included both academic and community practices from January 1, 2000, to December 31, 2023. The Comprehensive Longitudinal Investigation of Multiple Sclerosis at Brigham and Women's Hospital (CLIMB) registry is a prospective cohort that has followed MS patients with standardized clinical visits, including EDSS assessments, since 2000.[20] To construct the EHR-based MS cohort, we first identified all patients with at least one MS diagnosis code. Given the high sensitivity but low specificity of ICD codes for disease classification,[21]we then applied a knowledge graph-guided "weakly supervised" phenotyping algorithm[22] to improve diagnostic accuracy and reduce misclassification. Using a classification threshold calibrated to achieve 90% specificity (reaching 92% positive predictive value, PPV) through validation against a manually annotated subset of MS diagnosis, we identified 18,560 patients for downstream imputation of RRMS subtype and EDSS and causal analysis.[19]

Feature Set for Imputation and Confounder Adjustment

We included EHR features to achieve two objectives: (1) impute RRMS subtype and disability outcome, and (2) adjust for potential confounding in causal analyses.

For each patient, we extracted monthly counts of codified EHR data, including diagnosis codes (*e.g.*, International Classification of Disease [ICD]), procedure codes (*e.g.*, Current Procedural Terminology [CPT]), prescriptions (*e.g.*, RxNorm), and laboratory tests (*e.g.*, Logical Observation Identifiers Names and Codes [LOINC]).[23] We mapped all ICD codes to PheCodes[21] , consolidated CPT codes using the Clinical Classifications Software (CCS) for Services and Procedures, and grouped electronic prescription codes at the RxNorm ingredient level.[24] Clinical notes were processed using a natural language processing (NLP)[25,26] tool to generate monthly counts of medical terms, which were mapped to Concept Unique Identifiers (CUIs) according to the Unified Medical Language System (UMLS).[27] MS-relevant codes and medical terms were identified using multi-source clinical knowledge graphs.[22,28] To account for healthcare utilization, we quantified the total counts of codified EHR features. Counts of codified and narrative EHR features were log transformed.

To further adjust for confounding, we included standard demographic and clinical features commonly used in RCTs: *i.e.,* age at treatment initiation, sex (male vs female), race and ethnicity (non-Hispanic White vs otherwise), disease duration (defined as the time from the first recorded MS PheCode to treatment initiation, *i.e.,* index date) and prior MS-related medications.

Informative covariates were shown in **Supplementary Table 1**.

RRMS Subtype Imputation

MS subtypes in the CLIMB registry were recorded longitudinally at irregular intervals (*i.e.,* a median of three entries per patient). For patients with linked EHR data, registry-derived RRMS based on clinician documentation were used as the gold-standard RRMS status (*i.e.,* observed RRMS) at the corresponding time for training the subtype imputation model.

To account for changes in MS subtypes, models were trained to impute RRMS status (*i.e.,* imputed RRMS) using features (see Method - Feature Set) aggregated over consecutive two-year windows. RRMS classification models were developed using both standard supervised learning methods (*i.e.,* LASSO, XGBoost, Random Forest) and a semi-supervised approach (*i.e.,* LATTE, Label-efficient incident phenotyping)[29] which additionally leveraged surrogates and unlabeled EHR data from patients outside the registry. Final RRMS classifications were derived using a cross-validated ensemble of all models (*i.e.,* "Ensemble" method).

Imputation model performance was evaluated in the held-out set comprising 25% of registry patients excluded from training. For the RRMS subtype classification, we reported the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and PPV.

Clinician-Rated Disability Outcome Imputation

The clinician-rated disability measure based on EDSS ranges from 0 to 10 in 0.5 increment. As EDSS of 4.0 typically represents disability threshold significant enough to impact patient life, we operationally classified EDSS ≥4.0 as "moderate/severe disability," and <4.0 as "no/mild disability". Given that EDSS scores in the CLIMB registry were typically recorded every six months (*i.e.,* a median of four assessments per patient), we computed the median EDSS score and defined binary threshold (observed EDSS ≥4.0 vs <4.0) within consecutive six-month windows as the gold-standard clinician-rated disability labels.

To account for changes in EDSS scores and impute missing scores, models were trained to impute binary EDSS threshold (*i.e.,* imputed EDSS) using features (see Method - Feature Set) aggregated over consecutive 6-month windows. We trained models using data from a randomly selected subset of 500 patients with available EDSS scores. Binary EDSS classification models were developed using standard supervised learning methods (*i.e.,* LASSO, XGBoost, Random Forest) and the semi-supervised LATTE method. Predicted likelihood (*i.e.,* probability) for moderate or severe clinician-rated disability was derived using a cross-validated ensemble of all models (*i.e.,* "Ensemble" method).

Imputation model performance was evaluated in the held-out set comprising registry patients excluded from training by comparing predicted likelihood of moderate/severe disability to the gold-standard EDSS≥4.0 labels. For the binary EDSS classification, we reported AUC, sensitivity, specificity, and PPV.

Patient-Reported Disability Outcome Imputation

The patient-reported disability measure based on Patient Determined Disease Steps (PDDS) provides crucial patient perspectives and complements clinician-rated EDSS score.[30] Thus, we also imputed baseline PDDS scores (See methods in **Supplementary Materials S1.1)**.

Target Population

To illustrate a framework for scalable and robust RWE generation, we set out to compare two DMTs that are both considered highly effective for MS but lack head-to-head RCT evidence. Only one DMT (*i.e.,* OCR) is approved for PPMS and one DMT (*i.e.,* mitoxantrone, rarely

prescribed) has been approved for non-active SPMS. In contrast, multiple DMTs are approved for RRMS. Thus, we focused on two commonly prescribed DMTs for RRMS: OCR and NTZ.

The target population in this study were patients with the RRMS subtype at treatment initiation. Baseline RRMS status was determined using either registry-derived labels or imputed labels, with the latter classified using a threshold calibrated to achieve PPV≥0.90.

Inclusion criteria included: (1) treatment initialization between 2012 and 2020 to ensure sufficient pre-treatment and post-treatment follow-up, (2) no prior exposure to other high-effectiveness DMTs or chemotherapy agents (*e.g.*, Alemtuzumab, Cyclophosphamide, Cladribine, Mitoxantrone and Rituximab [RTX]), and (3) at least six months of available EHR data before and after the index date (see Treatment Exposure).

Treatment Exposure

For the *primary* causal analysis, we compared OCR and NTZ. For the *exploratory* causal analysis, we compared B-cell depletion therapy (including OCR and RTX) and NTZ (**Supplementary Material S1.2, Supplementary Figure 1**). Treatment initiation was defined as the first recorded instance of the corresponding medication code in the EHR and designated as the index date (*i.e.,* baseline). Patients were assigned to the target exposure group based on the first treatment received.

Treatment Outcome

The primary study outcome was the likelihood of developing moderate/severe disability (EDSS≥4.0) during two years after target treatment initiation, modeled as a continuous probability between 0 and 1. We included observed EDSS scores when available and imputed EDSS scores

when gold-standard labels are unavailable. To address temporal fluctuations in EDSS scores and noise in observed measurement or imputation, we first estimated the moderate/severe disability likelihood in each of the four consecutive 6-months periods following treatment initiation and then aggregated these estimates to derive a two-year cumulative likelihood. This approach provides a more robust representation of overall disability status during the two-year follow-up.

Statistical Analysis for Estimating Average Treatment Effect

We estimated the average treatment effect (ATE) using a doubly robust causal inference framework that combined propensity score modeling with calibration regression on imputed outcomes to minimize bias.

To adjust for potential confounding, the analysis included 254 features measured at the index date, including (1) standard demographic and clinical features (see Feature Set): *e.g.*, age, sex, race and ethnicity, and disease duration as well as imputed likelihood of binary baseline clinician-rated and patient-reported disability status during the two years preceding the index date; (2) EHR features relevant to "multiple sclerosis" or "disability", selected using multi-source clinical knowledge graphs [19,22] (*e.g.,* MS diagnosis codes, mentions of "magnetic resonance imaging" in clinical notes) and aggregated over the two years preceding the index date; (3) interaction terms between baseline disability status and other covariates to enable stratification by disability status. Covariate balance was assessed by comparing weighted covariate distributions between treatment exposure groups using tests of independence.

Propensity scores for treatment exposure and calibration regressions were estimated using penalized logistic regression, which automatically selected informative covariates while shrinking non-informative ones. We excluded patients with extreme propensity scores (≥ 0.99 or ≤ 0.01) since

they were highly unlikely to have received the alternative treatment, limiting the ability to infer counterfactual outcomes. We then applied augmented inverse probability weighting (AIPW) to integrate these components and compute the ATE.

To assess robustness, we conducted two sensitivity analyses to evaluate the consistency of ATE estimates when using gold-standard labels rather than imputed ones: (1) restricting to patients with registry-derived RRMS subtype; (2) restricting to patients with both registry-derived RRMS subtype and observed EDSS assessments within the two-year follow-up window.

RESULTS

RRMS Subtype and Disability Status Imputation

For imputing the RRMS subtype, LATTE achieved an AUC of 0.879 in the held-out test set, while the Ensemble method combining the machine learning methods marginally enhanced the prediction to AUC of 0.894 (**Table 1**). Using a pre-specified calibrated threshold of ≥0.63 for the RRMS likelihood (to reach PPV≥0.90), the ensemble model achieved a sensitivity of 0.808, specificity of 0.831 and PPV of 0.901 and was used to define a cohort with imputed RRMS subtype (**Supplementary Table 5**).

For imputing the clinician-rated binary disability status (*i.e.,* likelihood of EDSS≥4.0) either during pre-treatment baseline or treatment follow-up, LATTE achieved an AUC of 0.926 in the held-out test set, while the Ensemble method again slightly improved the prediction to AUC of 0.931 (**Table 1**). Using a representative calibrated threshold of 0.78 for the likelihood of EDSS≥4.0 (to reach PPV≥0.90) for reporting and compared against observed labels, the ensemble model, which achieved a sensitivity of 0.506, specificity of 0.983, and PPV of 0.902 (at the representative threshold), was used for defining baseline and post-treatment clinician-rated

disability (**Supplementary Table 6**). Unlike the RRMS imputation to establish a cohort with the RRMS subtype, disability imputation (likelihood value ranging between 0 and 1) did not require a cut-off calibration threshold.

Please see **Supplementary Material 1.2**, **Supplementary Table 2** and **Supplementary Table 7** for performance metrics of imputing the patient-reported binary disability status. Again, the ensemble method was used for defining baseline patient-reported disability.

Patient Profile in the Causal Analysis Cohort

For the causal analysis cohort comparing OCR vs NTZ, we excluded 113 out of the 861 eligible patients with imputed RRMS subtype due to extreme propensity scores, leaving a final cohort of 748 patients with imputed RRMS subtype receiving either treatment (543 OCR vs 205 NTZ) (**Figure 1**). Before covariate adjustment, treatment groups differed in baseline characteristics (**Supplementary Table 1**). Compared with OCR-treated patients at baseline, NTZ-treated patients had higher predicted pre-treatment disability measured by EDSS (OCR: 0.16 vs NTZ: 0.21; p=0.003), higher predicted pre-treatment disability measured by PDDS (OCR: 0.12 vs NTZ: 0.14; p=0.02), lower healthcare usage measured by EHR codes (log-counts OCR: 2.74 vs NTZ: 2.40; p<0.001) and fewer MS diagnosis codes (log-counts OCR: 2.41 vs NTZ: 2.03, p<0.001). These baseline differences suggest potential confounding by disease severity and healthcare utilization. After covariate adjustment by IPW, treatment groups were balanced on baseline characteristics (**Figure 2**).

Primary Causal Analysis

In the primary causal analysis using imputed RRMS subtype and imputed disability outcome (*i.e.,* likelihood of EDSS≥4.0 during two years after treatment initiation), OCR-treated patients had a lower two-year risk of moderate to severe disability (5.38%, 95% CI: 3.40%-7.36%) compared with NTZ-treated patients (11.25%, 95% CI: 6.79%-15.71%). The absolute risk difference was –5.87% (95% CI: –11.28% to –0.46%; p=0.033), favoring OCR (**Figure 3**).

Sensitivity Causal Analysis

Sensitivity analyses confirmed that ATE estimates remained consistent when using gold-standard labels (**Figure 3**). When analyzing 571 patients with a registry-derived RRMS subtype, OCR was associated with a lower two-year risk of moderate to severe disability (6.46%, 95% CI: 4.17%–8.74%) compared with NTZ (12.42%, 95% CI: 7.79%–17.04%), yielding a statistically significant risk difference of –5.96% (95% CI: –11.77% to –0.15%; p=0.045) favoring OCR. When analyzing 317 patients with both a registry-derived RRMS subtype and observed EDSS outcomes (*i.e.,* gold-standard observed labels only), OCR also showed a lower two-year disability risk (9.13%, 95% CI: 7.36%–10.9%) compared with NTZ (14.79%, 95% CI: 6.23%–23.35%), but the difference was not statistically significant (risk difference –5.66%; 95% CI: –14.15% to 2.84%; p=0.192), likely reflecting the smaller sample size and greater variability.

Exploratory Causal Analysis

During the study period, RTX was commonly used off label in RRMS. For exploratory causal analysis comparing B-Cell Depletion therapy (BCD: OCR or RTX) with NTZ, BCD was

associated with lower two-year disability risk than NTZ across all population definitions (**Supplementary Materials S1.2, Supplementary Figure 2-3, Supplementary Table 3-4**.)

DISCUSSION

This study aimed to strengthen RWE for head-to-head comparison of treatment effects on disability outcomes between OCR and NTZ in RRMS given the absence of RCT evidence. We integrated complementary registry and EHR data from a large healthcare system, applied ensemble machine learning methods to impute key missing patient information (*i.e.,* RRMS subtype, disability status), and developed a novel semi-supervised doubly robust approach with a calibration step to mitigate imputation bias. This new framework enables direct comparison of treatment effects on disability outcomes in target MS populations using large-scale real-world clinical data beyond registry patients. In the primary causal analysis as a clinically relevant illustration of the framework, OCR was associated with a significantly lower two-year risk of disability progression compared with NTZ in populations defined by both imputed and registry-derived RRMS subtype.

Findings in Context and Clinical Implications

To our knowledge, previous observational studies comparing OCR and NTZ relied exclusively on registry data, without extending analyses to patients outside registries or integrating EHR data that could provide complementary and informative patient characteristics often missing from disease-specific registries[31–36]. While most prior studies reported no significant differences between OCR and NTZ across various definitions of disability progression, Yamout et al. observed higher rates of *disability improvement* with NTZ compared with BCD therapy (*i.e.,* OCR and RTX) in RRMS.[32] .

Several key differences distinguish our study from the Yamout study beyond complementary data sources (registry-linked EHR cohort vs registry only). First, OCR- and RTX-treated patients in our study had distinct profiles, necessitating careful patient projection to the same target population before grouping both DMTs under BCD for comparison with NTZ, a step not addressed in the Yamout study. Second, the studies used different disability outcome definitions. The two-year risk of moderate or severe disability, aggregated across four consecutive 6-month windows in our study, is likely more stable than the 6-month confirmed disability change used in the Yamout study. Notably, Yamout et al. also reported no significant difference in disability progression between BCD and NTZ. Third, our doubly robust causal estimation approach is less vulnerable to model misspecification than the propensity scores matching method used in the Yamout study. Finally, differences in patient populations (United States vs Middle Eastern and North African) may further explain the divergent findings. Future RCTs are needed to provide the definitive evidence regarding whether OCR is superior to NTZ in delaying disability progression in target populations.

Inclusion of Rituximab with Ocrelizumab for Causal Analysis

An important and unperhaps surprising conclusion from this study is the need to project RTX- and OCR-treated patients onto a common target population for causal analysis, given the distinct patient profiles of these treatment groups. In MS research, DMTs with similar mechanisms of action are typically grouped together: *e.g.*, RTX and OCR under BCD therapy. However, their use in routine practice could differ substantially.

RTX is frequently prescribed off label, giving clinicians greater flexibility in dosing and frequency compared with OCR. In contrast, OCR administration is typically restricted by

insurance requirements to a fixed regimen (*i.e.,* 600 mg every six months) according to the drug approval. Clinically, RTX is more often administered in patients with more aggressive disease because of this flexibility for higher dose and frequency.

Prior literature supports our observation of differences between RTX and OCR groups[37–42]: RTX-treated patients tend to have longer disease duration, higher baseline disability scores, and more prior DMT exposure. Prior studies also reported higher relapse rates, greater likelihood of treatment discontinuation, increased hospitalization risk, and different safety profiles for RTX compared with OCR. These differences underscore the need for caution and care when grouping RTX and OCR under BCD therapy, as selection and indication biases may confound results.

To address this in our exploratory causal analysis comparing BCD (OCR or RTX) vs NTZ, we pre-defined a common target population and then projected patients receiving different treatments onto that shared covariate space. Specifically, RTX-treated patients in the exploratory analysis were mapped to the covariate distribution of the primary causal analysis cohort (OCR vs NTZ) (see **Supplementary Material S1.2**).

Future studies could conduct separate head-to-head comparisons of RTX vs NTZ and RTX vs OCR using methods that explicitly account for differences in patient profiles. Such analyses will help clarify whether observed heterogeneity reflects treatment selection patterns or true differences in treatment effectiveness.

Strengths

This study has several noteworthy strengths. First, by linking registry and EHR data, we expanded the target population beyond registry patients for causal analysis. Leveraging gold-standard registry labels, imputing key missing features (*i.e.,* RRMS subtype, baseline disability

status, and two-year disability outcome) from EHR data using ensemble machine learning more than doubled the sample size (*i.e.,* 2.36-fold, from 317 to 748 eligible patients), increasing statistical power and improving the generalizability beyond registry-only cohorts.

Second, our causal analysis incorporated a rich set of covariates selected through multi-source clinical knowledge graphs, including NLP-extracted concepts, alongside standard clinical and demographic features, patient-reported and clinician-rated baseline disability status, and healthcare utilization, many of which were not considered in prior studies.

Third, our novel semi-supervised doubly robust causal estimation framework, which combines machine learning-based imputation with a calibration step to mitigate bias, addresses major challenges in real-world data where key patient profiles and outcomes are often incomplete. This approach ensures reliable and interpretable treatment effect estimates in routine clinical settings. Importantly, this study demonstrates the feasibility of scaling comparative effectiveness research using real-world data while maintaining methodological rigor.

Finally, it provides methodologically robust and clinically relevant real-world evidence of comparative effectiveness (*i.e.,* OCR vs NTZ in RRMS) to bridge the gap left by the absence of head-to-head RCTs, offering practical guidance for MS treatment decisions and generating hypotheses for future validation through pragmatic trials. Taken together, the study framework establishes a foundation for comparing other MS therapies and potentially extending clinical investigations to other chronic conditions.

Limitations

Several limitations warrant consideration. First, although our imputation models for RRMS subtype and disability status performed well in held-out test sets, residual misclassification may

still occur and could bias treatment effect estimates. To mitigate this, we calibrated imputed RRMS subtype against gold-standard registry labels, ensuring that all registry-confirmed RRMS patients were retained in the target population. Assuming modest misclassification in the non-registry EHR population, any residual error is likely to increase variance rather than introduce substantial bias. For disability outcome imputation, we incorporated calibration regression within the causal analysis framework to further reduce misclassification and improve validity of treatment effect estimates.

Second, unmeasured confounding remains an inherent challenge in observational research. Despite adjusting for a comprehensive set of codified and narrative EHR variables, residual differences, such as disease severity or socioeconomic factors, may persist between treatment groups. Our imputation approach enabled adjustment for unmeasured baseline disability status. Similar strategy could extend to other partially observed measures available through linked registry data. While *individual* socioeconomic factors are often difficult to capture in EHR data, future studies could incorporate *neighborhood-level* socioeconomic indicators based on residence and census data.

Third, because OCR was approved more recently than NTZ, temporal shifts in clinical practice and treatment effect are possible. To minimize this impact, we narrowed the study window to 2012-2020. Future studies could further address by adjusting for or matching on calendar year of treatment initiation.

Conclusion

This study presents a robust framework for comparing treatment effects on disability outcomes across MS therapies that lack head-to-head RCT evidence. It offers a general roadmap

for conducting scalable, bias-mitigated comparative effectiveness research using real-world clinical data. The semi-supervised doubly robust approach, combining machine learning-based imputation with a critical calibration step, addresses incomplete information that has previously limited the effective use of EHR data for causal inference. As a clinically relevant demonstration to address knowledge gaps, we provide real-world evidence that OCR is associated with lower two-year risk of moderate or severe disability compared with NTZ in RRMS by expanding the target population and applying bias-mitigated causal inference methods.

ACKNOWLEDGEMENT

Data Access and Code Availability

Code for analysis and figures is available at <https://github.com/xialab2016/MS_OCRvsNTZ.git>. We will publicly disseminate anonymous summary-level registry data and EHR data. The rationale for not sharing patient-level data is that patient-level clinical data (either de-identified information or limited protected health information containing dates of clinical events or even if anonymous due to concern for re-identification) are universally subject to the rules and regulation of each healthcare system, which may only be affiliated with but are not the same as the primary academic institutions of the study investigators. Sharing of de-identified EHR data with qualified external researchers by each of the study performance site may be permissible only after the approval of the respective Institutional Review Boards (IRBs), regulatory oversight agents of the healthcare systems (that own the clinical data) as well as the appropriate Data Usage Agreements (DUA) between institutions.

Funding Statement

This study was funded by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award numbers R01 NS098023 (Z. Xia).

FIGURES

Figure 1. **Overview of the study design and the patient cohort for causal inference**. *EDSS,* Expanded Disability Status Scale; *EHR*, electronic health records; *KG*, knowledge graph; *LASSO,* Least Absolute Shrinkage and Selection Operator; *LATTE*, Label-efficient Incident phenoTyping; *NTZ*, natalizumab; *OCR*, ocrelizumab; *RF,* random forest; *RRMS,* relapsing-remitting multiple sclerosis; *XGBoost*, extreme gradient boosting.

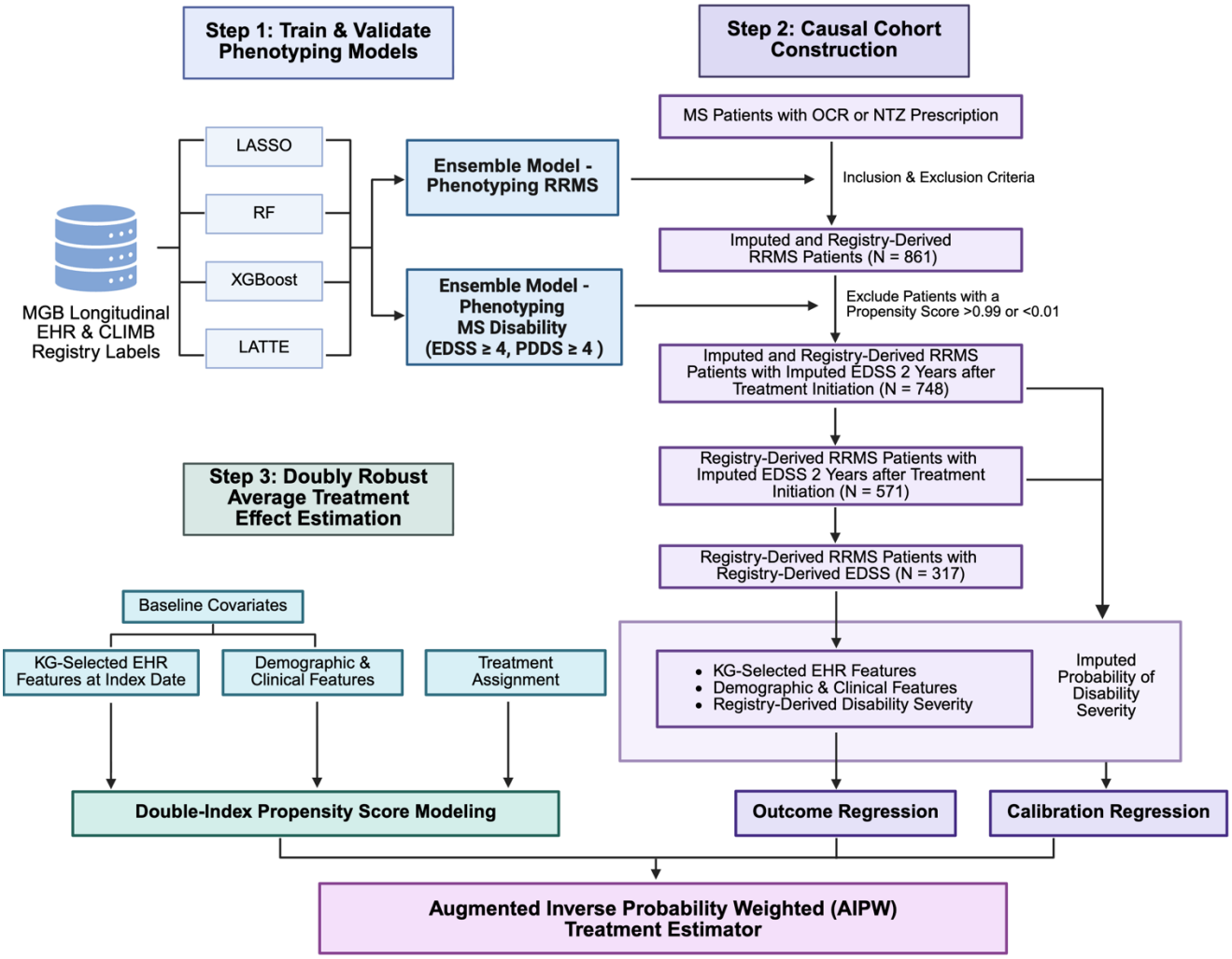


Figure 2. Covariate balancing. P-values before and after inverse probability weighting (IPW) are shown for all covariates with non-zero coefficients in either the propensity score or outcome regression models in the **primary causal analysis** comparing the 2-year disability between **OCR** versus **NTZ**. The vertical dashed line at p=0.05 indicates the threshold for nominal significance. P-values were derived from chi-squared tests for categorical variables or two-sample Kolmogorov–Smirnov tests for continuous variables. All covariates were marginally balanced (p>0.05) post-weighting.

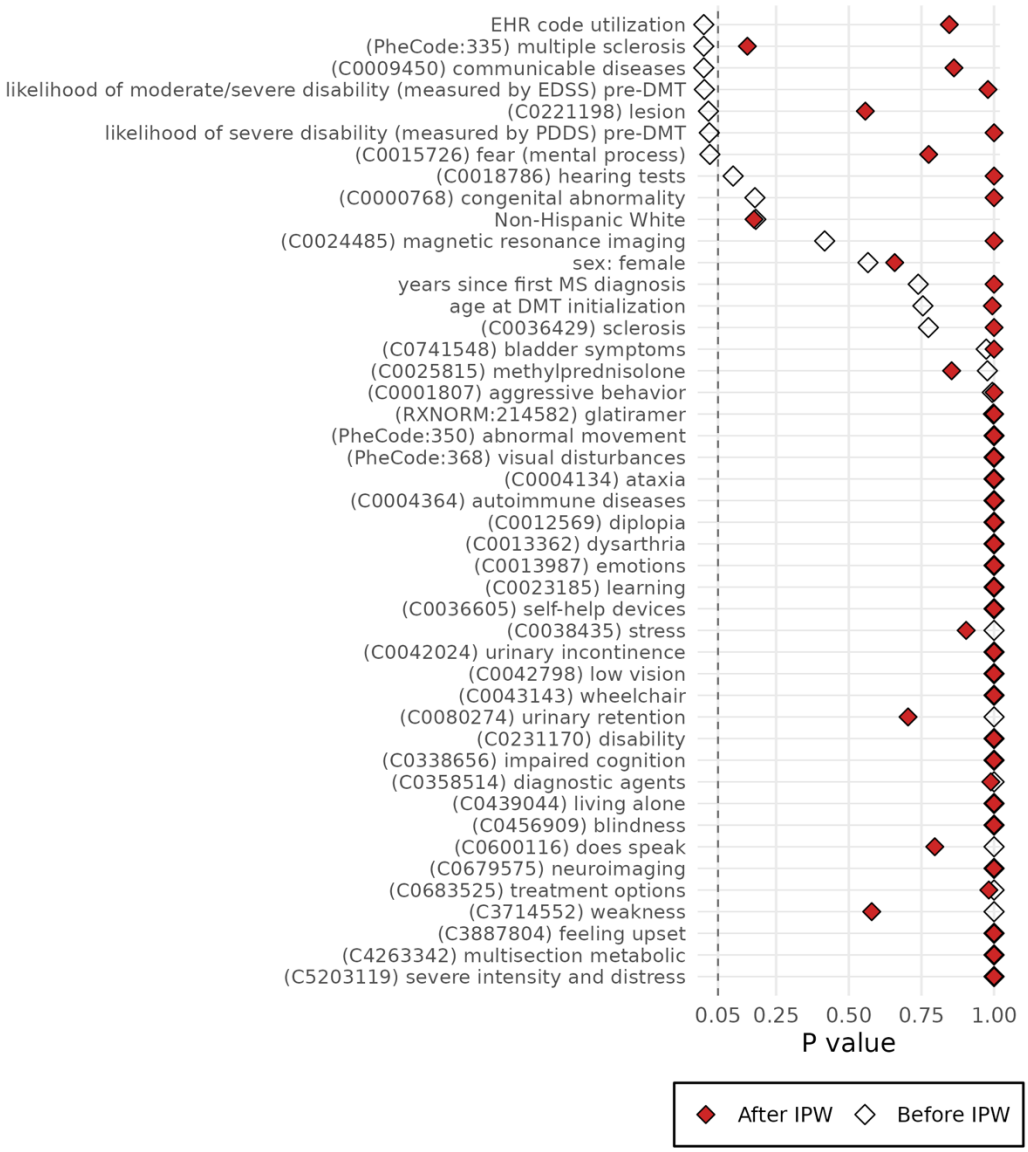
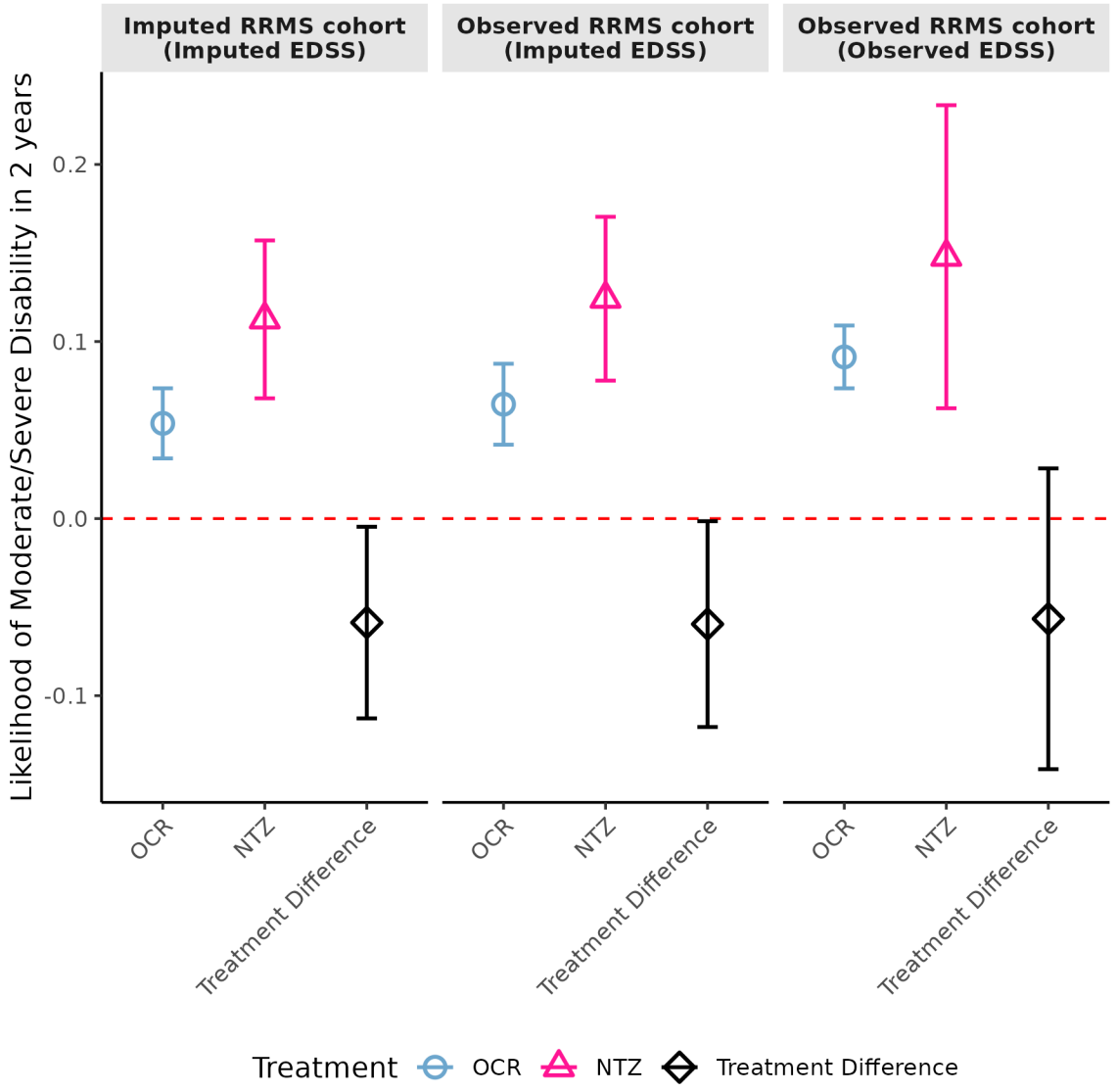


Figure 3. Average treatment effects of OCR vs NTZ. Estimated average treatment effects (markers) on the likelihood of moderate or severe disability (EDSS≥4.0) during the two years after treatment initiation are shown with pointwise 95% confidence intervals (vertical bars). OCR was associated with a lower two-year risk of moderate or severe disability across analyses, when using imputed EDSS in patients with imputed RRMS subtype, imputed EDSS in registry patients with observed RRMS subtype or observed EDSS in registry patients with observed RRMS subtype. While point estimates remained similar, the confidence intervals widened when using registry-derived subtype and disability status.



TABLES

Table 1. Imputation predictive performance. Area under the receiver operating characteristic curve (AUC) with 95% confidence intervals for imputing the **RRMS subtype** and for imputing clinician-rated moderate to severe disability (**EDSS≥4.0**) by the different machine learning methods is shown.

Imputed Outcome	Ensemble Method *	LATTE	LASSO	Random Forest	XGBoost	**Training Set:** patient number (total records)	**Test Set:** patient number (total records)
RRMS subtype	0.894 (0.879-0.909)	0.879 (0.862-0.895)	0.861 (0.843-0.879)	0.832 (0.810-0.854)	0.857 (0.839-0.875)	8,159 (17,955)	2,720 (6,000)
EDSS≥4.0	0.931 (0.928-0.935)	0.926 (0.922-0.930)	0.872 (0.867-0.878)	0.873 (0.868-0.879)	0.858 (0.852-0.863)	500 (2,560)	8,869 (49,451)

REFERENCES

1. Jakimovski D, Bittner S, Zivadinov R, et al. Multiple sclerosis. *Lancet* 2024; 403: 183–202.

2. Polman CH, O'Connor PW, Havrdova E, et al. A Randomized, Placebo-Controlled Trial of Natalizumab for Relapsing Multiple Sclerosis. *New Engl J Med* 2006; 354: 899–910.

3. Hauser SL, Waubant E, Arnold DL, et al. B-Cell Depletion with Rituximab in Relapsing–Remitting Multiple Sclerosis. *New Engl J Medicine* 2008; 358: 676–688.

4. Montalban X, Hauser SL, Kappos L, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *New Engl J Med* 2017; 376: 209–220.

5. Hauser SL, Bar-Or A, Weber MS, et al. Association of Higher Ocrelizumab Exposure With Reduced Disability Progression in Multiple Sclerosis. *Neurology - Neuroimmunol Neuroinflammation* 2023; 10: e200094.

6. Achtnichts L, Zecca C, Findling O, et al. Correlation of disability with quality of life in patients with multiple sclerosis treated with natalizumab: primary results and post hoc analysis of the TYSabri ImPROvement study (PROTYS). *BMJ Neurol Open* 2023; 5: e000304.

7. Rudick RA, Stuart WH, Calabresi PA, et al. Natalizumab plus Interferon Beta-1a for Relapsing Multiple Sclerosis. *New Engl J Med* 2006; 354: 911–923.

8. Hauser SL, Bar-Or A, Comi G, et al. Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *New Engl J Med* 2017; 376: 221–234.

9. Svenningsson A, Frisell T, Burman J, et al. Safety and efficacy of rituximab versus dimethyl fumarate in patients with relapsing-remitting multiple sclerosis or clinically isolated syndrome in Sweden: a rater-blinded, phase 3, randomised controlled trial. *Lancet Neurology* 2022; 21: 693–703.

10. Rudick RA, Lee J-C, Cutter GR, et al. Disability Progression in a Clinical Trial of Relapsing-Remitting Multiple Sclerosis: Eight-Year Follow-up. *Arch Neurol* 2010; 67: 1329–1335.

11. McGinley MP, Goldschmidt CH, Rae-Grant AD. Diagnosis and Treatment of Multiple Sclerosis. *Jama* 2021; 325: 765–779.

12. Sreenivasan AP, Vaivade A, Noui Y, et al. Conformal prediction enables disease course prediction and allows individualized diagnostic uncertainty in multiple sclerosis. *npj Digit Med* 2025; 8: 224.

13. Kurtzke JF. Rating neurologic impairment in multiple sclerosis An expanded disability status scale (EDSS). *Neurology* 1983; 33: 1444–1444.

14. Kalincik T, Cutter G, Spelman T, et al. Defining reliable disability outcomes in multiple sclerosis. *Brain* 2015; 138: 3287–3298.

15. Xia Z, Secor E, Chibnik LB, et al. Modeling Disease Severity in Multiple Sclerosis Using Electronic Health Records. *PLoS ONE* 2013; 8: e78927.

16. Liang L, Kim N, Hou J, et al. Temporal trends of multiple sclerosis disease activity: Electronic health records indicators. *Mult Scler Relat Dis* 2022; 57: 103333.

17. Zhang T, Goodman M, Zhu F, et al. Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. *Neurol - Neuroimmunol Neuroinflammation* 2020; 7: e864.

18. Hou J, Kim N, Cai T, et al. Comparison of Dimethyl Fumarate vs Fingolimod and Rituximab vs Natalizumab for Treatment of Multiple Sclerosis. *Jama Netw Open* 2021; 4: e2134627.

19. Gan Z, Zhu W, Tang W, et al. Knowledge Graph-Guided Identification of Multiple Sclerosis and Therapeutic Trend Analysis: Real-World Evidence from Two Large Healthcare Systems. *medRxiv* 2025; 2025.11.12.25340116.

20. Gauthier SA, Glanz BI, Mandel M, et al. A model for the comprehensive investigation of a chronic autoimmune disease: The multiple sclerosis CLIMB study. *Autoimmun Rev* 2006; 5: 532–536.

21. Zhang Y, Cai T, Yu S, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc* 2019; 14: 3426–3444.

22. Xiong X, Sweet SM, Liu M, et al. Knowledge-Driven Online Multimodal Automated Phenotyping System. *medRxiv* 2025; 2023.09.29.23296239.

23. McDonald CJ, Huff SM, Suico JG, et al. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clin Chem* 2003; 49: 624–633.

24. Nelson SJ, Zeng K, Kilbourne J, et al. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Méd Inform Assoc* 2011; 18: 441–448.

25. Yu S, Chakrabortty A, Liao KP, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *J Am Med Inform Assn* 2016; 24: e143–e149.

26. Yu S, Cai T, Cai T. NILE: Fast Natural Language Processing for Electronic Health Records. *arXiv*. Epub ahead of print 2013. DOI: 10.48550/arxiv.1311.6063.

27. Humphreys BL, Lindberg DA. The UMLS project: making the conceptual connection between users and the information they need. *Bull Méd Libr Assoc* 1993; 81: 170–7.

28. Hong C, Rush E, Liu M, et al. Clinical knowledge extraction via sparse embedding regression (KESER) with multi-center large scale electronic health record data. *npj Digit Med* 2021; 4: 151.

29. Wen J, Hou J, Bonzel C-L, et al. LATTE: Label-efficient incident phenotyping from longitudinal electronic health records. *Patterns* 2024; 5: 100906.

30. Foong YC, Merlo D, Gresle M, et al. The Patient-Determined Disease Steps scale is not interchangeable with the Expanded Disease Status Scale in mild to moderate multiple sclerosis. *Eur J Neurol* 2024; 31: e16046.

31. Signoriello E, Signori A, Lus G, et al. NEDA-3 achievement in early highly active relapsing remitting multiple sclerosis patients treated with Ocrelizumab or Natalizumab. *Mult Scler Relat Disord* 2024; 87: 105594.

32. Yamout B, Alroughani R, Mohamed SFA, et al. Comparative effectiveness of natalizumab and anti-CD20 monoclonal antibodies in relapsing-remitting multiple sclerosis: a real-world propensity-score matched study. *J Neurol, Neurosurg Psychiatry* 2025; jnnp-2024-335704.

33. Boz C, Ozakbas S, Terzi M, et al. The comparative effectiveness of fingolimod, natalizumab, and ocrelizumab in relapsing-remitting multiple sclerosis. *Neurol Sci* 2023; 44: 2121–2129.

34. Iaffaldano P, Lucisano G, Guerra T, et al. A comparison of natalizumab and ocrelizumab on disease progression in multiple sclerosis. *Ann Clin Transl Neurol* 2024; 11: 2008–2015.

35. Barbuti E, Castiello A, Pozzilli V, et al. Comparative effectiveness, safety and persistence of ocrelizumab versus natalizumab in multiple sclerosis: A real-world, multi-center, propensity score-matched study. *Neurotherapeutics* 2025; 22: e00537.

36. Pape K, Rolfes L, Steffen F, et al. Comparative effectiveness of natalizumab versus ocrelizumab in multiple sclerosis: a real-world propensity score–matched study. *Ther Adv Neurol Disord* 2022; 15: 17562864221142924.

37. Delgado SR, Faissner S, Linker RA, et al. Key characteristics of anti-CD20 monoclonal antibodies and clinical implications for multiple sclerosis treatment. *J Neurol* 2024; 271: 1515–1535.

38. Techa-angkoon P, Siritho S, Tisavipat N, et al. Current evidence of rituximab in the treatment of multiple sclerosis. *Mult Scler Relat Disord* 2023; 75: 104729.

39. Fernandes AA, Neves AL, Ferro D, et al. Clinical and analytical monitoring of patients with multiple sclerosis on anti-CD20 therapeutics: a real-world safety profile study. *Front Neurol* 2025; 15: 1500763.

40. Alcalá C, Quintanilla-Bordás C, Gascón F, et al. Effectiveness of rituximab vs. ocrelizumab for the treatment of primary progressive multiple sclerosis: a real-world observational study. *J Neurol* 2022; 269: 3676–3681.

41. Roos I, Hughes S, McDonnell G, et al. Rituximab vs Ocrelizumab in Relapsing-Remitting Multiple Sclerosis. *JAMA Neurol* 2023; 80: 789–797.

42. Cunha C, Matos S, Bernardes C, et al. Effectiveness of anti-CD20 therapies following natalizumab discontinuation: insights from a cohort study. *Mult Scler Relat Disord* 2025; 101: 106564.

SUPPLEMENTARY MATERIALS

S1. Supplementary Text

S1.1. Baseline Binary PDDS Imputation Method and Result

Baseline disability status is a potentially important confounder in causal analyses but frequently missing in real-world data. In addition to the clinician-rated disability measure based on the Expanded Disability Status Scale (EDSS) score, patient-reported disability measure based on the Patient Determined Disease Steps (PDDS) score provides crucial patient perspectives and complements EDSS. Because EDSS and PDDS are not interchangeable, inclusion of both disability measures at baseline for confounder adjustment more robustly mitigates potential bias. Thus, we imputed baseline disability status based on both EDSS (see Main Methods – Clinician-Rated Disability Imputation) and PDDS (see below).

PDDS is a validated patient-reported outcome measure that ranges from 0 to 8 in increments of 1. PDDS score of 4 indicates severe disability requiring full ambulatory support. We classified PDDS score ≥4 as severe disability and <4 as no, mild, or moderate disability. We computed the median PDDS score and defined the binary threshold (observed PDDS ≥4 vs <4) within consecutive six-month windows as the gold-standard patient-reported disability labels.

To impute the binary PDDS threshold (*i.e.,* imputed PDDS), we used the same features (see Main Methods - Feature Set, **Supplementary Table 1**) aggregated over consecutive 6-month windows and applied the same modeling framework as for EDSS imputation (see Main Methods – Clinician-Rated Disability Imputation). Specifically, we used an ensemble modeling approach combining standard supervised learning methods (*i.e.,* LASSO, XGBoost, Random Forest) with the semi-supervised LATTE method. We trained models using data from a randomly selected subset of 500 patients with available PDDS scores. Predicted likelihood (*i.e.,* probability) for

severe patient-reported disability was derived using a cross-validated ensemble of all models (*i.e.,* "Ensemble" method).

Imputation model performance was evaluated in the held-out set comprising registry patients excluded from training by comparing predicted likelihood of severe disability to the gold-standard PDDS≥4 labels. For the binary PDDS classification, we reported AUC*,* sensitivity, specificity, and PPV.

For imputing the patient-reported binary disability status (*i.e.,* likelihood of PDDS≥4) during pre-treatment baseline, LATTE achieved an AUC of 0.879 in the held-out test set, while the Ensemble method slightly improved the prediction to AUC of 0.899 (**Supplementary Table 2**). Using a representative calibrated threshold of 0.95 for the likelihood of PDDS≥4 (to reach PPV≥0.90) for reporting and compared against observed labels, the ensemble model, which achieved a sensitivity of 0.081, specificity of 0.999, and PPV of 0.923 (at the representative threshold), was used for defining baseline patient-reported disability (**Supplementary Table 7**). Unlike the RRMS imputation to establish a cohort with the RRMS subtype, disability imputation (likelihood value ranging between 0 and 1) did not require a cut-off calibration threshold.

S1.2. Exploratory Analysis: NTZ versus BCD with RTX Projected onto Target Population

S1.2.1. *Heterogeneity of Patients Receiving RTX vs Patients Receiving OCR or NTZ*

Ocrelizumab (OCR) is part of the B-cell depletion (BCD) therapy class, which also includes rituximab (RTX), ofatumumab (OFA) and ublituximab (UTX). During the study period (2012-2020, to ensure sufficient pre-treatment and post-treatment follow-up), the MGB EHR data contained no patients receiving OFA or UTX. RTX has been used as an off-label treatment of MS.

In the study dataset, RTX-treated and OCR-treated patients differed in their clinical and demographic characteristics (**Supplementary Table 3**). Compared with OCR-treated patients, RTX-treated patients had older age at disease-modifying therapy (DMT) initiation (mean 43.4 vs 42.2 years, $p<0.001$), a longer disease duration (6.6 vs 5.6 years, $p<0.001$), greater pre-treatment likelihood of EDSS ≥4 (0.25 vs 0.17, $p<0.001$), greater pre-treatment likelihood of PDDS≥4 (0.14 vs 0.13, $p<0.001$) and higher healthcare utilization.

S1.2.2. *Exploratory Causal Analysis Cohort Construction and Patient Profile: Method and Results*

To enable a fair comparison with natalizumab (NTZ), we projected the cohort to identify RTX-treated patients who were clinical profiles similar to those of the OCR- or NTZ-treated population in the primary causal analysis (**Supplementary Table 1**). To illustrate this projection, we displayed the distribution of RTX-treated patients relative to OCR- and NTZ-treated patients using a combined Partial Least Squares (PLS) regression and Uniform Manifold Approximation and Projection (UMAP) approaches for dimensionality reduction (**Supplementary Figure 3**).

Patient eligibility and treatment exposure criteria were consistent with those applied in the primary analysis (Main Method – Target Population).

Among 1,057 eligible patients, we identified 196 patients with imputed and registry-derived RRMS subtype who had received RTX. For the exploratory causal analysis, only 66 RTX-treated patients had covariate profiles comparable to those in the primary causal analysis cohort (748 total: 543 OCR vs 205 NTZ) based on projected similarity in the multivariate covariate space (**Supplementary Figure 1**). We then pooled those 66 RTX-treated patients with OCR-treated patients to form the BCD treatment exposure group, yielding a final exploratory causal analysis cohort of 814 RRMS patients (BCD: 609; NTZ: 205).

Before covariate adjustment, treatment groups differed in baseline characteristics (**Supplementary Table 4**). Compared with BCD-treated RRMS patients at baseline, NTZ-treated patients had higher predicted pre-treatment disability measured by EDSS (OCR: 0.21 vs NTZ: 0.18; p=0.028), lower healthcare usage measured by EHR code counts (log-counts OCR: 2.40 vs NTZ: 2.71; p=0.001), and fewer MS diagnosis codes (log-counts OCR 2.03 vs NTZ: 2.38; p<0.001) These baseline differences suggest potential confounding by disease severity and healthcare utilization. After covariate adjustment by IPW, treatment groups were balanced on baseline characteristics (**Supplementary Figure 2**).

S1.2.3. *Exploratory Causal Analysis: Method and Results*

We applied the same doubly robust causal inference framework, combining propensity score modeling with calibration regression on imputed outcomes to minimize bias (see Main Methods – Statistical Analysis for Estimating Average Treatment Effect).

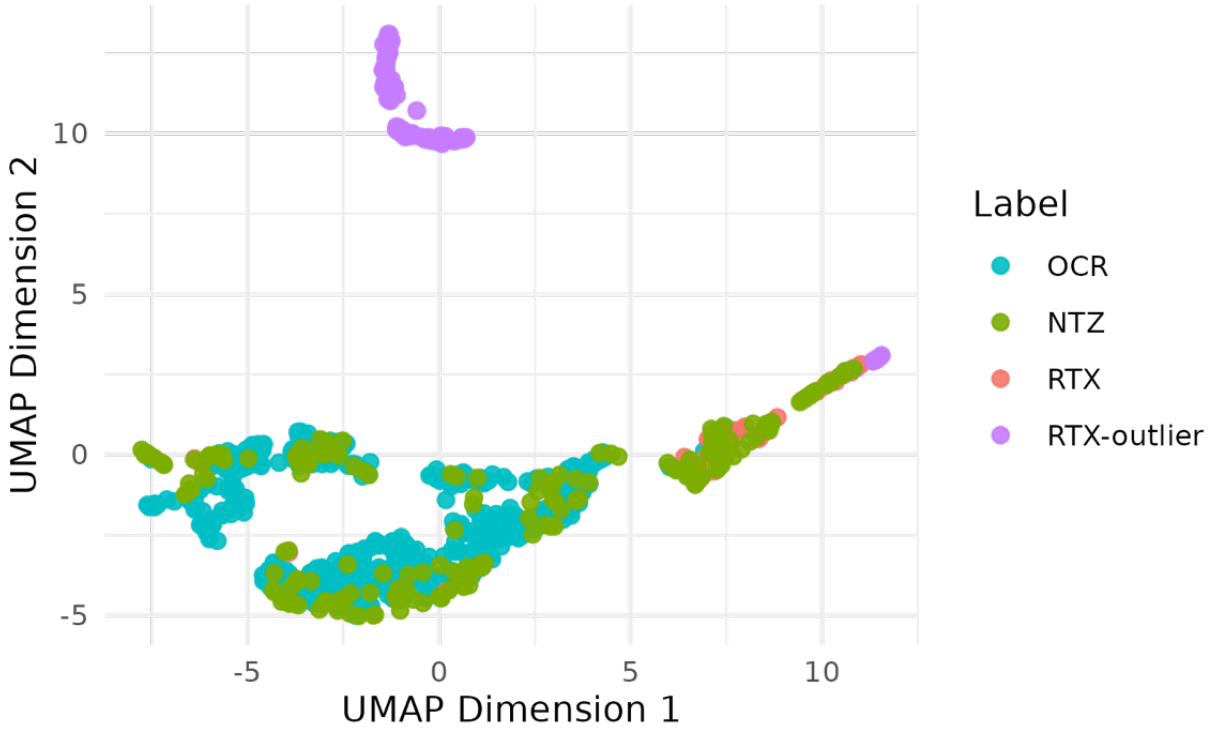
In the exploratory causal analysis using imputed RRMS subtype and imputed disability outcome (*i.e.,* likelihood of EDSS≥4.0 during two years after treatment initiation), BCD-treated patients had a lower two-year risk of moderate to severe disability (8.68%, 95% CI: 6.44%–10.93%) compared with NTZ-treated patients (14.53%, 95% CI: 10.60%–18.45%). The absolute risk difference was –5.84% (95% CI: –10.59 to –1.10; p=0.016), favoring BCD therapy (**Supplementary Figure 3**). Given the relatively modest sample size of the NTZ group, its ATE estimate had a wide 95% confidence interval and was less stable.

When analyzing 647 patients with a registry-derived RRMS subtype, BCD (8.30%, 95% CI: 6.09%–10.51%) was associated with a lower two-year risk of moderate to severe disability compared with NTZ (14.24%, 95% CI, 10.03%–18.45%). The absolute difference was –5.94% (95% CI: –11.26 to –0.63; p=0.028), favoring BCD.

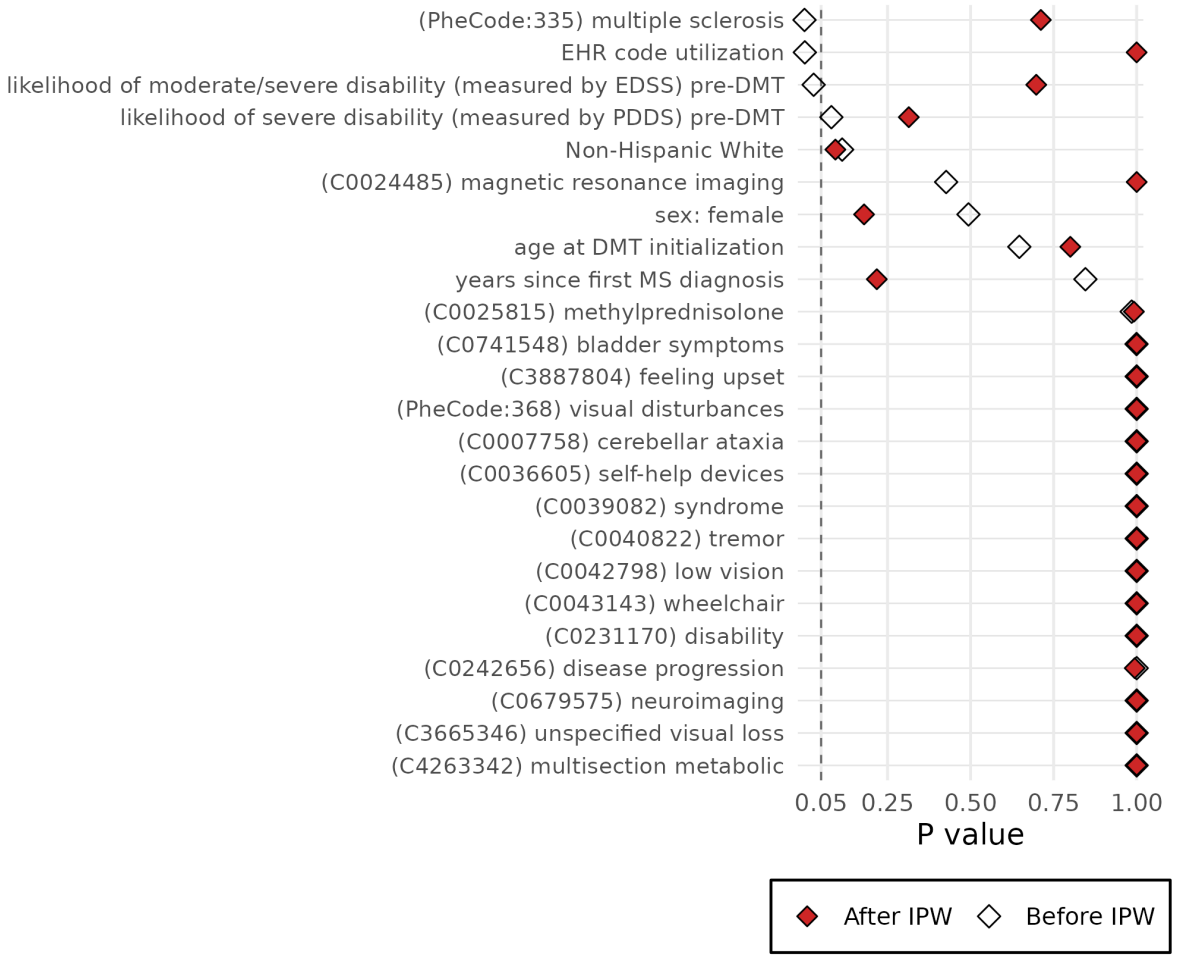
When analyzing 351 patients with both a registry-derived RRMS subtype and observed EDSS outcomes (*i.e.,* gold-standard observed labels only), BCD (8.80%, 95% CI: 6.37%–11.23%) also showed a lower two-year disability risk compared with NTZ (14.49%, 95% CI: 9.00%–19.98%). The risk difference was –5.69% (95% CI: –12.41% to 1.04%; p=0.097). This sensitivity analysis demonstrated a consistent trend, though it was not statistically significant, likely due to wider confidence intervals reflecting the smaller sample size and greater variability.

S2. Supplementary Figures

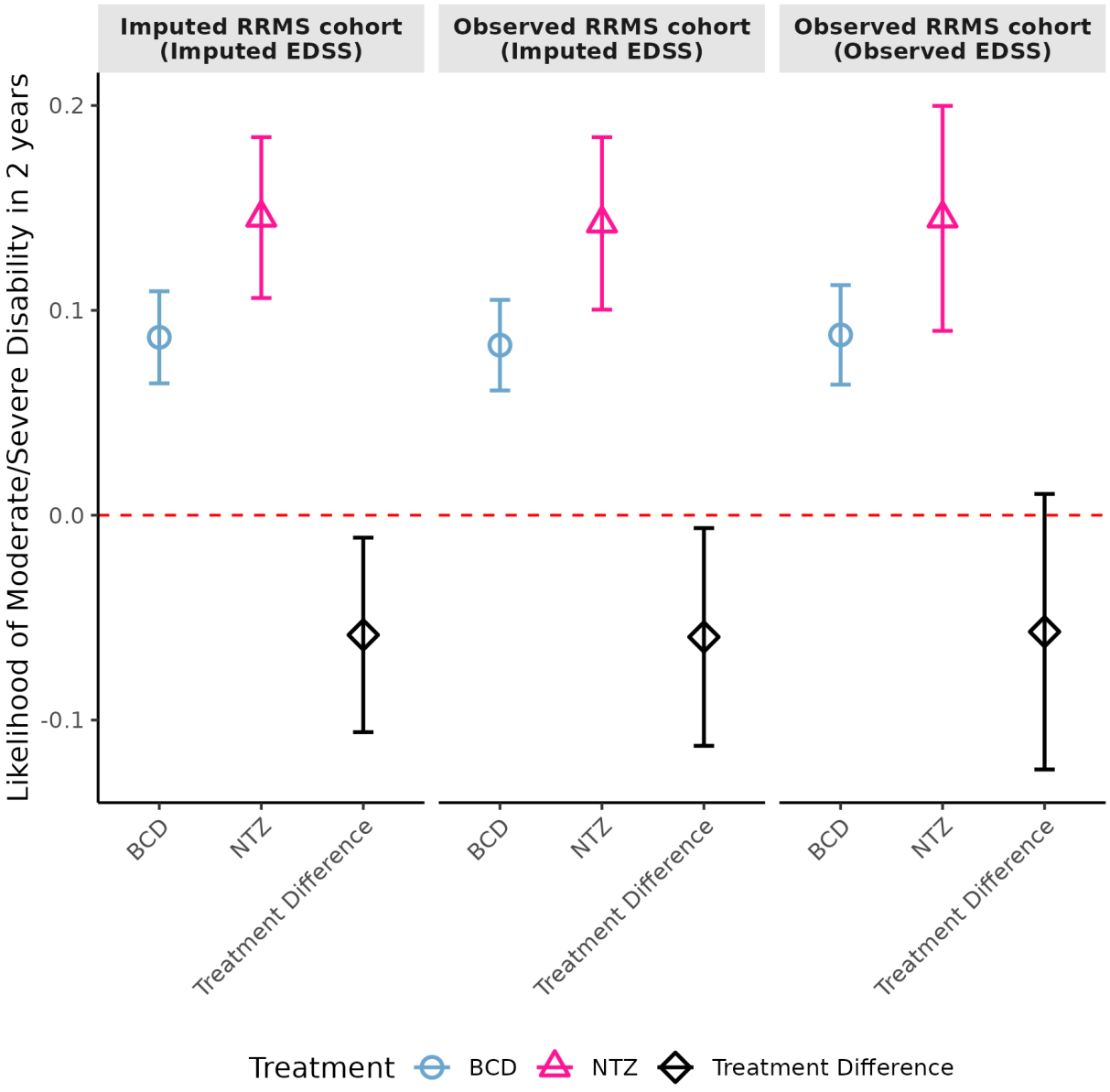
Supplementary Figure 1. Patient profile projection in multivariate space. Projection of patients receiving RTX onto the combined target patient population receiving OCR or NTZ. Patient embeddings were generated using a two-step dimensionality reduction approach that combined Partial Least Squares (PLS) regression and Uniform Manifold Approximation and Projection (UMAP). While OCR- and NTZ-treated patients (cyan and green) largely overlapped, most RTX-treated patients (red) formed a distinct cluster, with notable outliers, *i.e.,* RTX-outliers (purple). This pattern indicates that most RTX-treated patients generally did not share the same baseline profile as those in the primary causal analysis comparing OCR with NTZ.



Supplementary Figure 2. Covariate balancing. P-values before and after inverse probability weighting (IPW) are shown for all covariates with non-zero coefficients in either the propensity score or outcome regression models in the **exploratory causal analysis** comparing the 2-year disability between **BCD** versus **NTZ**. The vertical dashed line at p=0.05 indicates the threshold for nominal significance. P-values were derived from chi-squared tests for categorical variables or two-sample Kolmogorov–Smirnov tests for continuous variables. All covariates were marginally balanced (p>0.05) post-weighting.



Supplementary Figure 3. Average treatment effects of BCD vs NTZ. Estimated average treatment effects (markers) on the likelihood of moderate or severe disability (EDSS≥4.0) during the two years after treatment initiation are shown with pointwise 95% confidence intervals (vertical bars). **BCD therapy (OCR or RTX)** was associated with a lower two-year risk of moderate or severe disability across analyses, when using imputed EDSS in patients with imputed RRMS subtype, imputed EDSS in registry patients with observed RRMS subtype or observed EDSS in registry patients with observed RRMS subtype. While point estimates remained similar, the confidence intervals widened when using registry-derived subtype and disability status.



S3. Supplementary Tables

Supplementary Table 1. Covariate balance assessment for the *primary* causal analysis of comparing the two-year likelihood of moderate or severe disability between OCR versus NTZ. Features with non-zero coefficients in either the propensity score or outcome regression models are shown. Feature values are shown as mean (SD) or patient number (% total). Codified EHR feature counts and NLP-extracted narrative concept counts were log-transformed. Univariate associations were tested using either chi-squared tests for binary covariates or two-sample Kolmogorov–Smirnov tests for continuous covariates (unweighted and weighted by inverse propensity scores). All covariates were marginally balanced (p>0.05) post-weighting.

		Observed			**Inverse Probability Weighted**		
Feature	Overall	*OCR*	*NTZ*	*p-value*	*OCR*	*NTZ*	*p-value*
sex: female	573 (77 %)	413 (76 %)	160 (78 %)	0.566	425 (78 %)	163 (80 %)	0.658
non-Hispanic White	631 (84 %)	464 (85 %)	167 (81 %)	0.180	460 (85 %)	181 (88 %)	0.174
age at DMT initialization: years	42.22 (11.53)	42.87 (11.76)	40.49 (10.76)	0.755	42.47 (11.69)	41.08 (10.55)	0.994
years since first MS diagnosis	5.56 (5.31)	5.73 (5.42)	5.10 (4.97)	0.739	5.36 (5.28)	4.93 (5.11)	1.000
likelihood of moderate/severe disability (measured by EDSS) pre-DMT	0.17 (0.18)	0.16 (0.16)	0.21 (0.20)	0.003	0.17 (0.16)	0.19 (0.21)	0.979
likelihood of severe disability (measured by PDDS) pre-DMT	0.13 (0.07)	0.12 (0.07)	0.14 (0.08)	0.020	0.13 (0.08)	0.14 (0.13)	1.000
EHR code utilization	2.64 (0.85)	2.74 (0.76)	2.40 (0.99)	<0.001	2.61 (0.88)	2.77 (1.00)	0.846
(PheCode:335) multiple sclerosis	2.31 (0.71)	2.41 (0.59)	2.03 (0.89)	<0.001	2.27 (0.71)	2.41 (0.87)	0.151
(PheCode:350) abnormal movement	0.17 (0.49)	0.17 (0.50)	0.16 (0.45)	1.000	0.18 (0.50)	0.22 (0.52)	1.000

(PheCode:368) visual disturbances	0.10 (0.34)	0.11 (0.34)	0.10 (0.34)	1.000	0.11 (0.34)	0.08 (0.30)	1.000
(RXNORM:214582) glatiramer	0.24 (0.53)	0.20 (0.49)	0.32 (0.61)	0.998	0.18 (0.46)	0.32 (0.67)	1.000
(C0000768) congenital abnormality	0.72 (0.74)	0.76 (0.74)	0.62 (0.74)	0.177	0.73 (0.80)	0.79 (0.80)	1.000
(C0001807) aggressive behavior	0.09 (0.30)	0.07 (0.27)	0.14 (0.36)	0.994	0.10 (0.34)	0.16 (0.36)	1.000
(C0004134) ataxia	0.10 (0.35)	0.09 (0.32)	0.13 (0.42)	1.000	0.08 (0.30)	0.09 (0.34)	1.000
(C0004364) autoimmune diseases	0.09 (0.32)	0.10 (0.34)	0.07 (0.27)	1.000	0.09 (0.32)	0.06 (0.23)	1.000
(C0009450) communicable diseases	0.63 (0.70)	0.73 (0.70)	0.36 (0.62)	<0.001	0.71 (0.78)	0.56 (0.74)	0.862
(C0012569) diplopia	0.38 (0.65)	0.37 (0.64)	0.42 (0.68)	1.000	0.36 (0.62)	0.56 (0.84)	1.000
(C0013362) dysarthria	0.08 (0.29)	0.08 (0.27)	0.10 (0.33)	1.000	0.08 (0.27)	0.05 (0.24)	1.000
(C0013987) emotions	0.18 (0.44)	0.17 (0.44)	0.18 (0.43)	1.000	0.16 (0.42)	0.21 (0.44)	1.000
(C0015726) fear (mental process)	0.32 (0.55)	0.37 (0.57)	0.20 (0.48)	0.022	0.35 (0.57)	0.55 (0.79)	0.775
(C0018786) hearing tests	0.73 (0.71)	0.78 (0.72)	0.59 (0.69)	0.102	0.72 (0.71)	0.76 (0.76)	1.000
(C0023185) learning	0.15 (0.43)	0.13 (0.40)	0.21 (0.51)	1.000	0.12 (0.40)	0.26 (0.60)	1.000
(C0024485) magnetic resonance imaging	0.13 (0.34)	0.11 (0.31)	0.21 (0.39)	0.417	0.12 (0.32)	0.15 (0.33)	1.000
(C0025815) methylprednisolone	0.42 (0.69)	0.38 (0.62)	0.54 (0.83)	0.977	0.35 (0.60)	0.60 (0.84)	0.854
(C0036429) sclerosis	0.45 (0.64)	0.49 (0.68)	0.33 (0.54)	0.774	0.46 (0.65)	0.58 (0.77)	1.000
(C0036605) self-help devices	0.05 (0.22)	0.05 (0.21)	0.04 (0.23)	1.000	0.05 (0.23)	0.09 (0.37)	1.000

(C0038435) stress	0.34 (0.63)	0.36 (0.65)	0.30 (0.57)	1.000	0.33 (0.62)	0.50 (0.71)	0.904
(C0042024) urinary incontinence	0.16 (0.48)	0.17 (0.51)	0.14 (0.40)	1.000	0.19 (0.57)	0.18 (0.40)	1.000
(C0042798) low vision	0.06 (0.27)	0.06 (0.26)	0.07 (0.27)	1.000	0.05 (0.25)	0.06 (0.30)	1.000
(C0043143) wheelchair	0.06 (0.30)	0.06 (0.29)	0.08 (0.34)	1.000	0.08 (0.35)	0.14 (0.52)	1.000
(C0080274) urinary retention	0.28 (0.59)	0.27 (0.58)	0.29 (0.60)	1.000	0.28 (0.60)	0.54 (0.90)	0.704
(C0221198) lesion	1.53 (0.76)	1.61 (0.72)	1.32 (0.84)	0.017	1.52 (0.81)	1.63 (0.82)	0.557
(C0231170) disability	0.27 (0.59)	0.26 (0.60)	0.28 (0.56)	1.000	0.25 (0.58)	0.30 (0.60)	1.000
(C0338656) impaired cognition	0.14 (0.42)	0.12 (0.37)	0.20 (0.52)	1.000	0.12 (0.41)	0.24 (0.58)	1.000
(C0358514) diagnostic agents	0.25 (0.48)	0.27 (0.50)	0.21 (0.43)	1.000	0.31 (0.64)	0.38 (0.53)	0.989
(C0439044) living alone	0.15 (0.46)	0.15 (0.45)	0.16 (0.46)	1.000	0.15 (0.44)	0.19 (0.44)	1.000
(C0456909) blindness	0.10 (0.34)	0.11 (0.35)	0.09 (0.31)	1.000	0.13 (0.37)	0.12 (0.37)	1.000
(C0600116) does speak	0.24 (0.48)	0.23 (0.48)	0.26 (0.47)	1.000	0.24 (0.49)	0.44 (0.67)	0.796
(C0679575) neuroimaging	0.14 (0.36)	0.13 (0.35)	0.14 (0.38)	1.000	0.12 (0.34)	0.11 (0.31)	1.000
(C0683525) treatment options	0.50 (0.60)	0.51 (0.59)	0.50 (0.64)	1.000	0.47 (0.59)	0.65 (0.71)	0.982
(C0741548) bladder symptoms	0.11 (0.34)	0.09 (0.33)	0.16 (0.36)	0.973	0.08 (0.32)	0.10 (0.28)	1.000
(C3714552) weakness	0.82 (0.85)	0.81 (0.85)	0.84 (0.86)	1.000	0.77 (0.86)	1.02 (0.96)	0.579
(C3887804) feeling upset	0.09 (0.31)	0.07 (0.29)	0.13 (0.35)	1.000	0.08 (0.30)	0.13 (0.34)	1.000
(C4263342) multisection metabolic	0.08 (0.25)	0.07 (0.23)	0.08 (0.27)	1.000	0.08 (0.25)	0.17 (0.41)	1.000

(C5203119) intensity and distress	0.55 (0.77)	0.57 (0.79)	0.49 (0.70)	1.000	0.57 (0.80)	0.64 (0.94)	1.000

Supplementary Table 2. Imputation predictive performance. Area under the receiver operating characteristic curve (AUC) with 95% confidence intervals for imputing patient-reported severe disability (**PDDS ≥ 4**) by the different machine learning methods is shown.

Imputed Outcome	Ensemble Method *	LATTE	LASSO	Random Forest	XGBoost	**Training Set**: patient number (total records)	**Test Set**: patient number (total records)
PDDS≥4	0.899 (0.874-0.924)	0.879 (0.852-0.907)	0.867 (0.840-0.894)	0.754 (0.700-0.807)	0.828 (0.791-0.864)	500 (774)	1,982 (3,205)

Supplementary Table 3. Heterogeneity of Patients Receiving RTX vs Patients Receiving OCR or NTZ. Demographic and standard clinical characteristics of RTX-treated patients were compared with OCR- or NTZ-treated patients from the primary causal analysis (*i.e.*, the target population). Feature values are shown as mean (SD) or patient number (% total). P-values reflect statistical comparisons between groups using either chi-squared tests for binary covariates or two-sample Kolmogorov–Smirnov tests for continuous covariates.

	Observed		
Variable	Target population (OCR+NTZ)	RTX	p-value
sex: female	573 (77 %)	148 (76 %)	0.748
Non-Hispanic White	631 (84 %)	155 (79 %)	0.078
age at DMT initialization	42.22 (11.53)	43.41 (11.32)	<0.001
years since first MS diagnosis	5.56 (5.31)	6.59 (5.30)	<0.001
likelihood of moderate/severe disability (measured by EDSS) pre-DMT	0.17 (0.18)	0.25 (0.24)	<0.001
likelihood of severe disability (measured by PDDS) pre-DMT	0.13 (0.07)	0.14 (0.08)	<0.001
EHR code utilization	2.64 (0.85)	2.74 (0.85)	<0.001
(PheCode:335) multiple sclerosis	2.31 (0.71)	2.33 (0.73)	<0.001

Supplementary Table 4. Covariate balance assessment for the *exploratory* causal analysis of comparing the 2-year disability between BCD versus NTZ. Features with non-zero coefficients in either the propensity score or outcome regression models are shown. Feature values are shown as mean (SD) or patient number (% total). Codified EHR feature counts and NLP-extracted narrative concept counts were log-transformed. Univariate associations were tested using either chi-squared tests for binary covariates or two-sample Kolmogorov–Smirnov tests for continuous covariates (unweighted and weighted by inverse propensity scores). All covariates were marginally balanced (p>0.05) post-weighting.

		Observed			**Inverse Probability Weighted**		
Feature	Overall	*OCR*	*NTZ*	*p-value*	*OCR*	*NTZ*	*p-value*
sex: female	621 (76 %)	461 (76 %)	160 (78 %)	0.494	459 (75 %)	163 (80 %)	0.179
non Hispanic White	691 (85 %)	524 (86 %)	167 (81 %)	0.113	510 (84 %)	181 (88 %)	0.093
age at DMT initialization	42.37 (11.55)	43.00 (11.74)	40.49 (10.76)	0.647	42.18 (11.96)	41.08 (10.55)	0.800
years since first MS diagnosis	5.61 (5.34)	5.78 (5.45)	5.10 (4.97)	0.846	6.17 (5.33)	4.93 (5.11)	0.218
likelihood of moderate/severe disability (measured by EDSS) pre-DMT	0.18 (0.19)	0.18 (0.19)	0.21 (0.20)	0.028	0.24 (0.23)	0.19 (0.21)	0.698
likelihood of severe disability (measured by PDDS) pre-DMT	0.13 (0.07)	0.13 (0.07)	0.14 (0.08)	0.081	0.15 (0.08)	0.14 (0.13)	0.314
EHR code utilization	2.63 (0.85)	2.71 (0.78)	2.40 (0.99)	0.001	2.55 (0.91)	2.77 (1.00)	1.000
(PheCode:335) multiple sclerosis	2.29 (0.72)	2.38 (0.63)	2.03 (0.89)	<0.001	2.21 (0.77)	2.41 (0.87)	0.712
(PheCode:368) visual disturbances	0.10 (0.33)	0.10 (0.33)	0.10 (0.34)	1.000	0.08 (0.27)	0.08 (0.30)	1.000
(C0007758) cerebellar ataxia	0.10 (0.32)	0.08 (0.28)	0.13 (0.41)	1.000	0.15 (0.37)	0.09 (0.33)	1.000
(C0024485) magnetic resonance imaging	0.13 (0.34)	0.11 (0.32)	0.21 (0.39)	0.426	0.13 (0.34)	0.15 (0.33)	1.000
(C0025815) methylprednisolone	0.43 (0.69)	0.39 (0.64)	0.54 (0.83)	0.985	0.51 (0.73)	0.60 (0.84)	0.992
(C0036605) self-help devices	0.05 (0.23)	0.05 (0.23)	0.04 (0.23)	1.000	0.07 (0.36)	0.09 (0.37)	1.000

(C0039082) syndrome	0.15 (0.44)	0.16 (0.47)	0.11 (0.33)	1.000	0.10 (0.37)	0.17 (0.40)	1.000
(C0040822) tremor	0.25 (0.52)	0.26 (0.53)	0.24 (0.48)	1.000	0.22 (0.50)	0.30 (0.58)	1.000
(C0042798) low vision	0.06 (0.27)	0.06 (0.27)	0.07 (0.27)	1.000	0.09 (0.37)	0.06 (0.30)	1.000
(C0043143) wheelchair	0.07 (0.32)	0.07 (0.31)	0.08 (0.34)	1.000	0.11 (0.37)	0.14 (0.52)	1.000
(C0231170) disability	0.27 (0.59)	0.26 (0.60)	0.28 (0.56)	1.000	0.22 (0.56)	0.30 (0.60)	1.000
(C0242656) disease progression	0.13 (0.32)	0.13 (0.30)	0.15 (0.38)	1.000	0.08 (0.26)	0.22 (0.43)	0.994
(C0679575) neuroimaging	0.13 (0.35)	0.13 (0.34)	0.14 (0.38)	1.000	0.10 (0.29)	0.11 (0.31)	1.000
(C0741548) bladder symptoms	0.11 (0.35)	0.10 (0.35)	0.16 (0.36)	1.000	0.14 (0.36)	0.10 (0.28)	1.000
(C3665346) unspecified visual loss	0.10 (0.36)	0.10 (0.38)	0.08 (0.30)	1.000	0.06 (0.29)	0.08 (0.32)	1.000
(C3887804) feeling upset	0.08 (0.30)	0.07 (0.28)	0.13 (0.35)	1.000	0.07 (0.26)	0.13 (0.34)	1.000
(C4263342) multisection metabolic	0.07 (0.24)	0.07 (0.23)	0.08 (0.27)	1.000	0.09 (0.24)	0.17 (0.41)	1.000

Supplementary Table 5. Imputation predictive performance: Sensitivity, positive predictive value (PPV) and negative predictive value (NPV) for imputing the **RRMS** subtype by the different machine learning methods using a *pre-specified* calibrated threshold of ≥0.63 (to reach PPV≥0.90 for the **Ensemble** method) is shown.

Method \ Metric	Specificity	Sensitivity	PPV	NPV	Calibration Method
LASSO	0.849	0.698	0.898	0.597	0.63
Random Forest	0.794	0.740	0.872	0.616	0.63
XGBoost	0.760	0.787	0.861	0.652	0.63
LATTE	0.874	0.696	0.913	0.602	0.63
Ensemble	0.831	0.808	0.901	0.695	0.63

Supplementary Table 6. Imputation predictive performance: sensitivity, positive predictive value (PPV) and negative predictive value (NPV) for imputing clinician-rated moderate to severe disability (**EDSS≥4.0**) using a *representative* threshold of 0.5 (for reporting performance metrics) by the different machine learning methods or using a calibrated threshold of ≥0.78 (to reach PPV≥0.90 for the **Ensemble** method) is shown. Unlike the RRMS imputation to establish a cohort with the RRMS subtype, disability imputation (likelihood value ranging between 0 and 1) did not require a cut-off calibration threshold. We applied calibration only for the final ensemble method, while individual machine learning method was not calibrated. Of note, since the **predicted likelihood of EDDS ≥4.0** using LATTE are all smaller than 0.78, some performance metrics such as specificity, PPV or NPV may be NA.

Method \ Metric	Specificity	Sensitivity	PPV	NPV	Calibration Threshold
LASSO	0.961	0.509	0.806	0.860	0.5
Random Forest	0.981	0.412	0.874	0.840	0.5
XGBoost	0.945	0.504	0.746	0.857	0.5
LATTE	NA	1.000	0.242	NA	0.5
Ensemble	0.956	0.670	0.828	0.901	0.5

Method \ Metric	Specificity	Sensitivity	PPV	NPV	Calibration Threshold
LASSO	0.986	0.322	0.879	0.820	0.78
Random Forest	0.999	0.061	0.965	0.770	0.78
XGBoost	0.966	0.424	0.800	0.840	0.78
LATTE	1.000	0.000	NA	0.758	0.78
Ensemble	0.983	0.506	0.902	0.862	0.78

Supplementary Table 7. Imputation predictive performance: Sensitivity, positive predictive value (PPV) and negative predictive value (NPV) for imputing patient-reported severe disability (**PDDS ≥ 4**) using a *representative* calibration threshold of 0.5 (for reporting performance metrics) by the different machine learning methods or using a calibrated threshold of ≥0.95 (to reach PPV≥0.90 for the **Ensemble** method) is shown. Unlike the RRMS imputation to establish a cohort with the RRMS subtype, disability imputation (likelihood value ranging between 0 and 1) did not require a cut-off calibration threshold. We applied calibration only for the final ensemble method, while individual machine learning method was not calibrated. Of note, since the **predicted likelihood of PDDS ≥4** using random forest and LATTE are smaller than 0.95, some performance metrics such as specificity, PPV, or NPV may be NA.

Method \ Metric	Specificity	Sensitivity	PPV	NPV	Calibration Threshold
LASSO	0.951	0.372	0.579	0.893	0.5
Random Forest	0.977	0.318	0.712	0.888	0.5
XGBoost	0.916	0.486	0.511	0.908	0.5
LATTE	NA	1.000	0.153	NA	0.5
Ensemble	0.906	0.655	0.557	0.936	0.5

Method \ Metric	Specificity	Sensitivity	PPV	NPV	Calibration Threshold
LASSO	0.998	0.115	0.895	0.862	0.95
Random Forest	1.000	0.000	NA	0.847	0.95
XGBoost	0.973	0.243	0.621	0.877	0.95
LATTE	1.000	0.000	NA	0.847	0.95
Ensemble	0.999	0.081	0.923	0.857	0.95