Comparison of B-cell depletion versus natalizumab for treatment of multiple sclerosis: A semi-supervised causal analysis

Dominic DiSanto[1] *, Weijing Tang, PhD[2] *, Wen Zhu, MD[3], Michele Morris, BS[4], Shyam Visweswaran, MD, PhD[4], Tianxi Cai, ScD[1,5] #, Zongqi Xia, MD, PhD[3] #

1. Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
2. Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA
3. Department of Neurology, University of Pittsburgh, Pittsburgh, PA
4. Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
5. Department of Biomedical Informatics, Harvard Medical School, Boston, MA

*** Share co-first authorship**

 # Share co-senior authorship

Abstract

Background: B-cell depletion (BCD) therapies (*e.g.,* ocrelizumab, ofatumumab, rituximab) and natalizumab (NTZ) are highly effective disease-modifying therapies (DMTs) for multiple sclerosis (MS). However, no randomized clinical trial and only limited observational studies compared the two DMT classes.

Objective: We compared BCD and NTZ in managing MS patient-reported disability progression using registry-linked electronic healthcare record (EHR) data.

Methods: The study population of an EHR cohort of MS patients included a subset enrolled in a clinic-based MS registry that provided gold-standard outcome labels. To estimate average treatment effects, we applied a doubly-robust semi-supervised approach to analyze all (not only registry) patients and comprehensively adjusted for confounders that included not only *a priori* standard features but also knowledge graph-derived EHR features. While gold-standard disability outcomes were available in registry patients, we imputed the baseline pre-treatment and post-treatment disability status for non-registry patients. We categorized patient-reported disability progression status as “sustained worsening”, “sustained improvement”, or “no sustained change” based on 3 or more observations or imputations of Patient Determined Disease Steps (PDDS) scores within 3 years after target treatment initiation as the primary endpoint.

Results: In this MS cohort (n=1,738, Age=46±13 years, Non-Hispanic White=86.71%), there was no significant difference between BCD (n=1,245, 71.63%) and NTZ (n=495, 28.37%) in mitigating sustained worsening (ATE=0.012, 95% CI [-0.123, 0.102], p=.803) or promoting sustained improvement (ATE=-0.0823, 95% CI [-0.194, 0.022], p=.172) of patient-reported disability. Sensitivity analyses using a 2-year window after treatment initiation confirmed no difference in sustained worsening (ATE=-0.0141, 95% CI [-0.107, 0.080], p=.895) or sustained improvement (ATE=-0.0713, 95% CI [-0.263, 0.026], p=.245) between BCD and NTZ. In power

analysis, the semi-supervised approach increased statistical power compared to the standard approach of using gold-standard data alone.

Conclusion: This real-world comparative effectiveness analysis based on a novel doubly-robust semi-supervised approach found no difference between BCD and NTZ in managing MS disability progression.

Key Messages

- The scarcity of patient-reported disability outcomes in routine clinical care hinders analysis based on real-world patient-reported outcomes, while evaluation of sustained disability accumulation requires long-term follow-up in clinical trials.
- Using a large registry-linked electronic healthcare record cohort and a novel semi-supervised, doubly-robust method, we conducted a causal inference study to compare sustained change in patient-reported disability in people with MS.
- The semi-supervised approach effectively leverages additional data from patients without observed outcome information and increases the statistical power of the comparative effectiveness study while retaining robustness properties in the causal analysis.
- There was no statistically significant difference between B-cell depletion therapy and natalizumab in sustained patient-reported disability outcomes up to 3-years after treatment initiation.

Introduction

Multiple sclerosis (MS) is characterized by inflammatory demyelination and progressive neurodegeneration in the central nervous system. Both neuroinflammation (*i.e.,* relapses) and neurodegeneration contribute to neurological disability accumulation in people with MS (pwMS). Growing disease-modifying therapies (DMTs) have improved MS outcomes by reducing relapses and delaying disability worsening. Among commonly prescribed DMTs in the United States (US), B-cell depletion (BCD) therapies (*e.g.,* ocrelizumab, ofatumumab, rituximab) and natalizumab (NTZ) are associated with reduced disability worsening in pwMS.[1–12] However, there is no randomized clinical trial (RCT) and limited robust real-world evidence directly comparing the effectiveness of these pairs of DMTs in managing disability.[13–24]

The clinician-assessed Expanded Disability Status Scale (EDSS) is the most recognized disability measure in MS. While commonly used in RCTs, EDSS is impractical for routine clinical monitoring or in most research settings due to high costs (*e.g.,* need for trained personnel) and high patient burden (*e.g.,* time commitment). In contrast, the Patient-Determined Disease Steps (PDDS) scale is a commonly used patient-reported outcome (PRO) for assessing MS disability. While PDDS complements EDSS, PDDS is significantly easier to administer and has been increasingly adopted by research registries and even routine clinical practice for assessing patient-reported disability.[25]

While registries collect outcomes (*e.g.* PDDS) unavailable in routine patient care, electronic health records (EHR) provide longitudinal clinical data collected during routine care. These complementary data sources have enabled EHR-linked MS registry studies to model disease severity,[26] systematically identify comorbidities associated with MS severity,[27] predict future

relapse,[28] evaluate long-term temporal trends in relapse rate,[29] and adjust for confounding biases when comparing DMTs for relapse reduction.[30] However, registry participants may not represent the broader patient population, potentially limiting the generalizability of findings based solely on registries, even when enriched by linked EHR data. Advances in statistical methods have introduced doubly-robust estimation for average treatment effects (ATEs) in semi-supervised settings.[31,32] These methods are applicable when few patients have observed outcome measurements. Deploying semi-supervised methods, we can leverage both registry-enrolled patients with outcome data and crucially non-registry patients without outcome data in EHR for estimating treatment effects. This study employs a novel doubly-robust semi-supervised approach to assess the comparative effectiveness of two highly effective DMT classes (BCD vs NTZ) in managing patient-reported sustained disability outcomes in pwMS.

Methods

Ethics Approval

The University of Pittsburgh Institutional Review Board approved the study protocols (STUDY19080007, STUDY21030127). All registry participants provided written informed consent. EHR research protocol was deemed exempt.

Data Source

We used inpatient and outpatient EHR data from the University of Pittsburgh Medical Center (UPMC). Codified data (available from January 1, 2004 to December 31, 2022) contained demographic (*e.g.*, age, gender, race and ethnicity) and clinical information, including daily counts of codes indicating diagnoses (*e.g.,* International Classification of Disease [ICD]), procedures (*e.g.,* current procedural terminology [CPT]), prescriptions (*e.g.,* RxNorm), and laboratory tests (*e.g.,* Logical Observation Identifiers Names and Codes [LOINC]). To organize codified data, we mapped all ICD codes to PheCodes,[33] consolidated CPT procedure codes using the Clinical Classifications Software (CCS) for Services and Procedures, and grouped electronic prescription codes to the RxNorm ingredient level (see **Supplementary Materials-EHR Data Pre-processing**). Narrative data (free-text clinical narrative available from January 1, 2011 to December 31, 2022) underwent a previously validated natural language processing pipeline according to unified medical language systems to generate concept unique identifiers (CUIs).[34,35]

We used EHR-linked registry data of a subset of patients enrolled in a clinic-based MS cohort, Prospective Investigation of Multiple Sclerosis in the Three Rivers Region (PROMOTE, Pittsburgh, PA), between 2017 and 2023. Registry data included longitudinal outcomes (*i.e.,*

PDDS) as well as past and present DMT records, which were integrated with prescriptions from EHR to assign treatment arms.

Cohort Derivation

The cohort follow-up window (2014-2022) strikes a balance between having sufficient sample size and mitigating temporal shift in clinical practice. We used the Knowledge-Driven Online Multimodal Automated Phenotyping (KOMAP) algorithm[36] to identify patients with an MS diagnosis in the EHR. Among 17,827 potentially eligible patients, 3,039 pwMS received their first target treatment (BCD or NTZ) between January 1, 2014, and December 31, 2022. We excluded 993 patients who had received other highly effective DMTs (*e.g.,* alemtuzumab, cladribine, mitoxantrone) or who received target treatment prior to January 1, 2014, as well as 76 patients who ever received chemotherapies (*e.g.,* cyclophosphamide, methotrexate). We further excluded 217 patients who switched to another DMT after target treatment initiation. Finally, we excluded 15 patients with missing demographic or clinical features. This final cohort included 1,738 MS patients, including 667 registry-enrolled participants.

Treatment Assignment

We compared two highly effective DMT mechanistic classes: BCD (*i.e.,* ocrelizumab, ofatumumab, rituximab) and NTZ. While rituximab is not officially approved for MS, it has been widely used in clinical practice. Ublituximab was approved after data extraction. We assigned patients in the BCD or NTZ treatment arm, if they first received either target treatment between January 1, 2014 and December 31, 2022. The date of first receiving BCD or NTZ was the target treatment initiation date.

Disability Outcomes

To assess treatment effectiveness, we used sustained changes in disability based on PDDS[37–39]. As a validated scale of patient-reported disability in MS, PDDS scores range from 0 to 8, with higher scores indicating greater disability (of gait impairment). Due to sparse observations of observed PDDS scores of 8 in the dataset (<1%), we combined PDDS scores of 7 and 8 into a single category for analysis.

First, we categorized patients into one of four groups: “sustained worsening” (≥1 PDDS score increase following treatment initiation, sustained for at least 6 months), “sustained improvement” (similarly defined but for ≥1 PDDS score decrease), “no sustained change”, and “insufficient information”. Specifically, adjudication of *sustained* changes required at least three PDDS scores, separated by at least 6 months between two consecutive scores, within 3 years after treatment initiation. Sustained worsening or improvement required two sequential PDDS scores that were either greater or lesser, respectively, compared to the first observed PDDS score after treatment initiation. When a patient experienced both sustained worsening and sustained improvement within the study period, we considered the first occurrence of sustained change as the outcome (*i.e.,* worsening or improvement, whichever occurred earlier). We refer to patients with “insufficient information” as “***unlabeled outcome***” or as having “***no labeled outcome***”, and the other three categories as having “***labeled***” outcomes. Labeled outcome data are available only in a subset of registry-enrolled participants.

Next, we assessed sustained worsening and sustained improvement separately. When evaluating sustained worsening, we merged the other two labeled categories of patients who did not experience sustained worsening as “no sustained worsening”: *i.e.,* patients with "sustained improvement" and "no sustained change". When evaluating sustained improvement, we likewise

merged the other two labeled categories as “no sustained improvement”: *i.e.,* patients with "sustained worsening" and "no sustained change".

Confounders

To control potential confounding, we adjusted for baseline features based on standard clinical trial practice, healthcare utilization, and knowledge-derived features from EHR. We obtained the covariates primarily from EHR and additionally from registry data, which provided a complementary source for registry-enrolled patients.

Standard clinical trial features included demographics (*i.e.,* age at target treatment initiation, self-reported gender, race and ethnicity), disease duration (*i.e.,* from the first MS PheCode to target treatment initiation), follow-up duration (*i.e.,* from the first clinical encounter to treatment initiation), prior DMT use duration (*i.e.,* from the earliest record of prior DMT use to treatment initiation; setting as 0 for those without any prior DMT).

As baseline disability status at treatment initiation was infrequently available, we calculated a baseline PDDS risk covariate for all patients using a PDDS imputation model and each patient’s available clinical history prior to target treatment initiation. PDDS imputation was developed using a separate dataset of PDDS scores unused in this causal analysis. See **Supplementary Materials-PDDS Imputation Model, Supplementary Tables 1a-b, 2; Supplementary Figures 1a-b**) for details on the PDDS imputation method and prediction performance.

Healthcare utilization features indicate not only overall health status but also potentially missing information in the fragmented US healthcare landscape. To capture *total* baseline healthcare utilization, we included a covariate of the counts of all observed EHR (codified and narrative)

features during distinct clinical encounters that occurred prior to target treatment initiation. Further, we included a covariate of the counts of distinct clinical encounters with an MS PheCode (based on distinct code-date pairings, *i.e.,* the number of unique dates with MS PheCode).

To select EHR features informative of MS and disability status while reducing feature dimensionality, we used an online narrative and codified feature search engine (ONCE), powered by pre-trained multi-source knowledge graph representation learning of EHR concepts.[36] See detailed methods in **Supplementary Materials-Knowledge Graph-Derived Feature Selection** and the list of included EHR features in **Supplementary Tables 3a-d**. To address feature sparsity, we excluded rare features with <10% frequency in the cohort. We used normalized aggregated counts of diagnoses (PheCode), procedures (CCS codes), prescriptions (RxNorm), laboratory tests (LOINC), and clinical narrative features (CUIs) before treatment initiation. Normalization was achieved by dividing the total occurrence of each feature by the total health care utilization.

Statistical Analysis

To conduct causal analyses, we employed a doubly-robust, semi-supervised estimation method.[40,41] To emulate RCT using observational data, it was crucial to adjust for confounders associated with both treatment assignment (*i.e.,* BCD vs NTZ) and outcome (*i.e.,* sustained change in PDDS). With confounder adjustment, the doubly-robust method yields valid estimates despite misspecification in either the propensity score model (*i.e.,* treatment assignment given baseline confounders) or the outcome model (*i.e.,* outcome given baseline confounders). In contrast to the standard doubly-robust methods that rely solely on labeled data (*i.e.,* observed outcome), this semi-supervised method incorporates both labeled and no labeled outcome

observations to improve efficiency in estimating comparative treatment effects. An overview of the study design is presented in **Figure 1**.

Figure 1 - Study Design

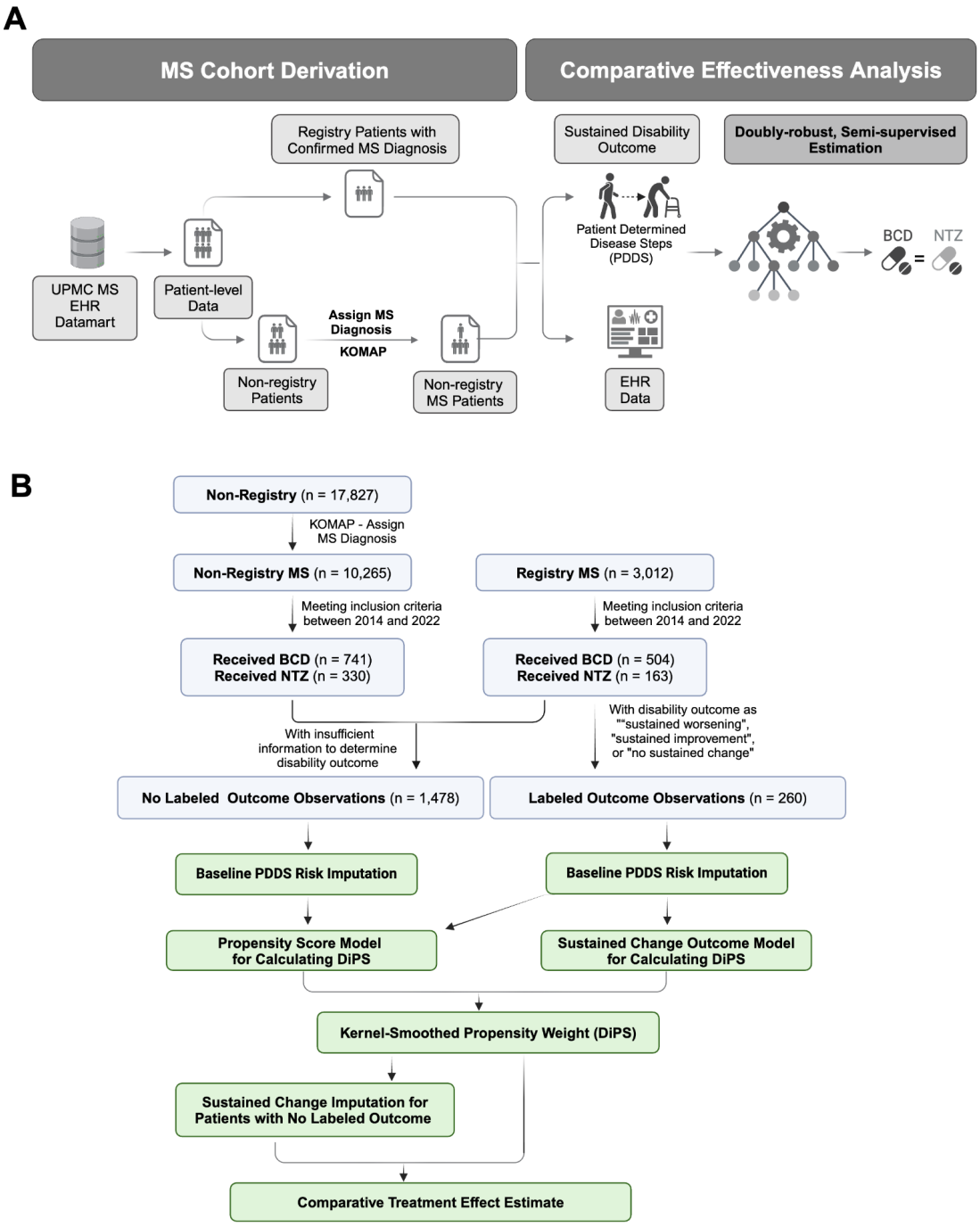

Figure 1. Study design. (A) Schematic overview. (B) Flowchart of cohort derivation steps (boxes with blue outline) and the modeling steps in the causal analysis (green boxes). The "Non-Registry - MS" panel represents patients with a KOMAP algorithm-phenotyped MS diagnosis who were not enrolled in the clinic-based MS registry. The "Registry - MS" panel represents patients with neurologist-confirmed MS diagnosis who were enrolled in the clinic-based MS registry (with linked EHR data). Sample sizes for the "Received" DMT categories correspond to patients with a first record of BCD or NTZ between January 1, 2014 and December 31, 2022. See Methods for details on the cohort derivation. "Labeled outcome observations" correspond to patients with "sustained worsening", "sustained improvement", or "no sustained change" in PDDS score, while "no labeled outcome observations" correspond to patients with "insufficient information" for assessing sustained changes.
Abbreviations: BCD, B-Cell Depletion therapy. DMT, Disease Modifying Therapy. EHR, Electronic Health Record. NTZ, natalizumab. PDDS, Patient Determined Disease Steps.

The semi-supervised estimation method consists of three steps. First, we fit initial propensity and outcome regression adaptive Least Absolute Shrinkage and Selection Operator (LASSO) models to calculate a modified version of a treatment propensity weight, specifically a kernel-smoothed propensity score, *i.e.* the **double-index propensity score (DiPS)**. Second, we developed an outcome imputation model that predicts sustained-change (in PDDS) outcomes for data with no labeled outcome, which were then incorporated into the final doubly-robust estimation. This outcome imputation model can use additional, surrogate information available after target treatment initiation. Finally, we estimate the treatment effect using a **doubly-robust inverse propensity weighting** (IPW) estimator, using the DiPS as balancing weights and incorporating imputed outcomes from the second step.[40,41] This doubly-robust IPW estimator ensures consistency under correct specification of either an initial propensity score model or initial outcome model. Additionally, it is robust to misspecification of the outcome imputation model and generates consistent estimates as if the true outcomes were observed.

In the second step above, the outcome imputation model of sustained change used updated covariates including post-target treatment initiation information. This imputation model included the same set of expert-defined, codified, and narrative features included in the initial outcome model, with codified and narrative features now aggregated through 1-year after target

treatment initiation. We then included three additional covariates post-target treatment features, two related to target treatment group and a third year-1 PDDS risk covariate. We included a binary covariate of the target treatment group and a weight of the inverse DiPS of a patient's assigned target treatment group. For patients with observed PDDS scores within 1 year post-treatment initiation, the year-1 PDDS risk was the mean of observed PDDS scores. For patients without PDDS scores within the same window, the year-1 PDDS risk was the predicted PDDS score using an independently developed PDDS imputation model using codified and narrative feature data through 1-year post-target treatment. We then trained a LASSO logistic regression model of sustained change using labeled observations. We then used this model to impute the sustained change outcome when there was no labeled outcome.

The final comparative treatment effect estimation between BCD and NTZ was based on DiPS-weighted outcomes on all patients, used the observed outcome for labeled data and imputed outcomes for observations with no labeled outcome. We employed a perturbation resampling procedure to construct confidence intervals for estimates and calculated two-sided p-values through confidence interval inversion.[41,42] We considered p-values below 0.05 as statistically significant after applying a Holm correction. Further, we conducted a power evaluation based on a normal approximation, using the standard error estimated from perturbation analysis. We compared the results and efficiency of the semi-supervised analysis with an observed-case analysis (*i.e.,* using only labeled data) using the final, doubly-robust IPW estimator. See details in **Supplementary Materials-Detailed Causal Analysis**.

Sensitivity Analyses

The pre-processed data for the main analyses used all available pre-treatment data to calculate pre-treatment features, trimmed rare EHR features with less than 10% frequency, chose the 6-

months window between post-treatment PDDS scores to ascertain sustainment of disability change, and used knowledge graph for initial feature screening. To validate the main findings, we performed a series of sensitivity analyses by altering each of these parameters. Specifically, we repeated the analyses using 6-month and 12-month pre-treatment periods, applying a less stringent 5% frequency threshold of EHR features, categorizing sustained-change outcome status with a 3-month window between PDDS scores, and using all EHR features rather than only knowledge graph-derived features. Finally, we repeated the main and sensitivity analyses to sustained changes within 2 years (rather than 3 years) after treatment initiation.

Data Availability

The code for analysis and figures is available at <https://github.com/domdisanto/BCD_NTZ_SemiSupervisedCausal>. De-identified data are available upon request to the corresponding author and with permission from the participating institutions.

Results

Patient Characteristics

When assessing the cohort characteristics (**Table 1**), registry and non-registry patients are comparable *except* for healthcare utilization, where registry-enrolled patients demonstrate higher mean total baseline healthcare utilization. In the overall study cohort, 1,245 (71.63%) patients received BCD and 1,478 (85.04%) did not have labeled sustained-change outcomes. **Figure 2** displays the relative average or proportion of characteristics by treatment class and by sustained change outcome category as compared to the overall mean. When compared to patients in the NTZ group, patients in the BCD group had longer disease duration (5.62 vs 4.33 years), follow-up duration (9.82 vs 8.05 years), higher mean total baseline healthcare utilization (1,561 vs 900 distinct codes and CUIs), higher baseline PDDS risk (2.13 vs 1.88), and fewer women (68.35% vs 80.93%). Patients with sustained disability improvement had lower baseline and year-1 PDDS risk and were more often women. Additional summaries, stratified by outcome category and target treatment group, are available in **Supplementary Tables 4a-c**. Baseline covariates are similarly compared after balancing by DiPS weights in **Supplementary Tables 5a-b**, demonstrating appropriate covariate balance by target treatment class.

Table 1 - Cohort Characteristics

Clinical or Demographic Variable	Overall (n=1,738)	Registry (n=667)	Non-Registry (n=1,071)	P-Value
Age at target treatment initiation, years, mean (SD)	45.98 (13.02)	44.51 (12.77)	46.89 (13.10)	<0.001
Self-reported gender: n (%)				0.678
Men	488 (28.08%)	183 (27.44%)	305 (28.48%)	
Women	1,250 (71.92%)	484 (72.56%)	766 (71.52%)	

Race and ethnicity: n (%)				0.982
White, Non-Hispanic	1,507 (86.71%)	579 (86.81%)	928 (86.65%)	
Other race and ethnic groups	231 (13.29%)	88 (13.19%)	143 (13.35%)	
Disease duration: years, mean (SD)	5.26 (4.77)	5.17 (5.00)	5.31 (4.62)	0.063
Follow-up duration: years, mean (SD)	9.32 (4.99)	9.85 (4.91)	8.98 (5.01)	<0.001
Target treatment class*: n (%)				<0.001
BCD	1,245 (71.63%)	504 (75.56%)	741 (69.19%)	
Ocrelizumab	*799 (45.97%)*	*263 (39.43%)*	*536 (50.05%)*	
Ofatumumab	*38 (2.19%)*	*1 (0.15%)*	*37 (3.45%)*	
Rituximab	*408 (23.48%)*	*240 (35.98%)*	*168 (15.69%)*	
NTZ	493 (28.37%)	163 (24.44%)	330 (30.81%)	
Baseline total healthcare utilization: number of codes, mean (SD)	1374.14 (1504.90)	1474.55 (1529.81)	1311.60 (1486.47)	<0.001
Baseline PDDS risk: mean (SD)	2.06 (0.71)	2.07 (0.73)	2.05 (0.70)	0.571
Year-1 PDDS risk: mean (SD)	2.10 (0.91)	2.04 (0.87)	2.13 (0.92)	*0.008*
Post-treatment disability outcome, n (%)				*<0.001*
No sustained change	73 (4.20%)	73 (10.94%)	0 (0.00%)	
Sustained improvement	80 (4.60%)	80 (11.99%)	0 (0.00%)	
Sustained worsening	107 (6.16%)	107 (16.04%)	0 (0.00%)	
No Labeled Outcome	1,478 (85.04%)	407 (61.02%)	1,071 (100.00%)	

Table 1. Cohort characteristics with mean (standard deviation) presented for continuous variables and count (percentage) presented for categorical variables. Baseline total healthcare utilization was the count of all EHR features observed during distinct clinical encounters occurring prior to target treatment initiation. Baseline PDDS risk was derived as the predicted mean PDDS as time of target treatment initiation, using independently fitted PDDS imputation models. Year-1 PDDS risk was calculated as the average of observed PDDS scores within 1 year post-treatment initiation for patients with any available PDDS data and otherwise imputed by the same, independent PDDS imputation models using additional covariate information through 1 year post-treatment initiation. P-values are reported from chi-square tests for categorical variables (race-

ethnicity, gender, target treatment category, and disability outcome) and Mann-Whitney U-tests for the remaining, continuous variables, comparing Registry and Non-Registry observations.
*Specific BCD DMT's sample size and total column percentage reported.
Abbreviations: BCD, B-cell depletion. PDDS, Patient Determined Disease Steps.

Figure 2 - Cohort Characteristics by Treatment Class and Outcome Category

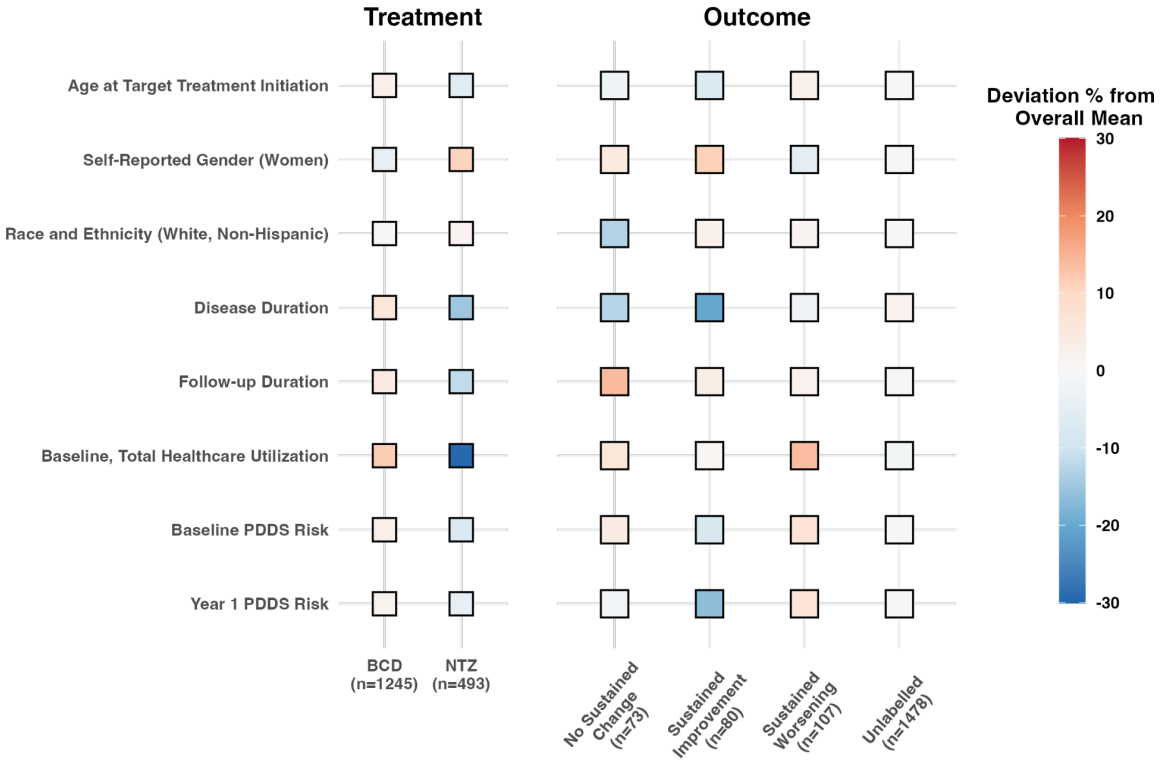

Figure 2. Cohort characteristics by treatment class and outcome category. Each square reports the difference between the column-wise subgroup mean and the overall cohort mean for the variable listed in each row. Red squares indicate values for a sub-group being greater than the overall mean, while blue squares indicate values being less than the overall mean. The left panel summarizes differences across the four sustained change categories. The right panel summarizes differences across treatment assignments. For categorical variables (Women, White, Non-Hispanic race/ethnicity), the mean corresponds to the percentage of the category.
Abbreviations: mo, month. PDDS, Patient Determined Disease Steps. BCD, B-Cell Depleting drugs. NTZ, Natalizumab.

Comparative Effectiveness Analysis

The main comparative effectiveness analysis results are presented in **Table 2**. Higher PDDS scores indicate greater patient-reported disability. With NTZ as the reference treatment class, a positive ATE for the sustained worsening outcome would indicate BCD as *less* effective than NTZ in reducing patient-reported disability worsening. Conversely, a positive ATE for the sustained improvement outcome would indicate BCD as *more* effective than NTZ in promoting patient-reported disability improvement. We observed no statistically significant difference between BCD and NTZ for either sustained worsening or sustained improvement during the 3-year post-treatment evaluation.

Table 2 - Treatment Effect Estimates

Outcome	ATE	Std. Err	95% CI	P-Value
Sustained Worsening	0.012	0.054	(-0.123, 0.102)	0.803
Sustained Improvement	-0.082	0.054	(-0.194, 0.022)	0.172

Table 2. Comparative treatment effect with NTZ as the reference treatment class. Visualization of the results is presented in **Supplementary Figure 2**.
Abbreviations: ATE, Average Treatment Effect; CI, Confidence Interval. Std. Err, Standard Error.

Features associated with treatment assignment and sustained-change in disability outcome informing the final, doubly-robust IPW estimator in the main analysis are shown in **Figure 3**. The same propensity score model and thus the informative features in the DiPS were used in analysis of both sustained-change outcomes. In the treatment model, men (as self-reported gender), higher baseline total healthcare utilization, mental disorders (PheCode: 306) and older age at target treatment initiation were most associated with a higher probability of receiving BCD than NTZ. Features associated with increased risk of sustained disability worsening included muscle spasms (PheCode 350.1), dysuria (PheCode 599.3), distress (C0231303), and

leukoencephalopathy (C0270612). Few features were selected in models of sustained disability improvement with the largest coefficients observed for treatment with NTZ and metabolic diagnosis procedures (C4263342).

Figure 3 - Coefficients in Treatment and Outcome Models for DiPS Calculation

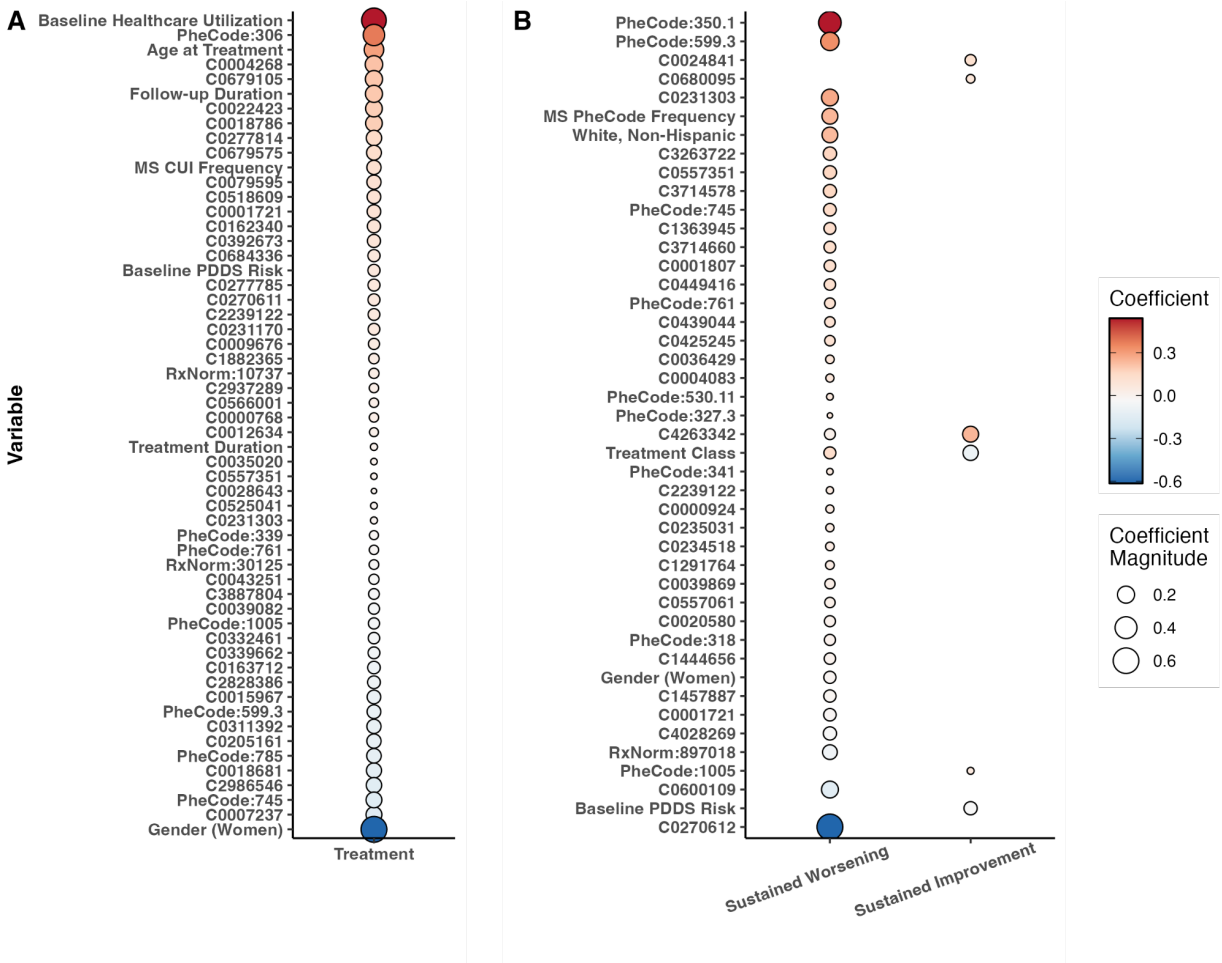

Figure 3. Coefficients in the initial propensity score model (**A**) and two initial outcome models (**B**) fitted to inform the kernel-smoothed propensity score estimator. The same treatment model was used for final causal analysis of each outcome. These models collectively informed the final inverse probability weight that was used to weight/balance for the final ATE calculation, but no balancing/weighting was used to calculate the coefficients shown here. Only variables with non-zero coefficients are included in each panel. Dot size represents the magnitude of the coefficient, while color corresponds to both effect size and direction. In Panel A, covariates with larger (positive) coefficients are associated with greater probability of receiving BCD than NTZ while the

inverse is for smaller coefficients. Detailed descriptions of CUI, PheCode, and RxNorm codes are included in **Supplementary Tables 3a-d**. All covariates were standardized (to mean 0, unit variance) prior to model fitting.
Abbreviations: CUI, Concept Unique Identifier.

Sensitivity Analyses and Power Evaluation

Sensitivity analyses reducing the pre-treatment periods, lowering the frequency threshold of EHR features, reducing the interval between PDDS scores in the sustained-change outcome definition, and using all available (rather than only knowledge graph-derived) EHR features produced treatment effect estimates and confidence intervals similar to the primary results (**Supplementary Figures 3a-b**, **Supplementary Tables 6a-b**), as did analyses using the 2-year post-treatment evaluation window (**Supplementary Table 7, Supplementary Figure 2**).

To quantify the advantage of the semi-supervised method leveraging observations with no labeled outcome, we compared the main semi-supervised analysis to an observed-case analysis (*i.e.,* using labeled data only) in the final ATE estimator (**Supplementary Tables 8a-b**). Fixing power at 80%, we reported the minimum detectable effect size for the risk difference in sustained improvement and sustained worsening between BCD and NTZ. Notably, the semi-supervised estimate is powered to detect a risk difference half the effect size of the observed-case estimate, indicating a meaningful gain in power when using the semi-supervised estimator.

Discussion

To fill knowledge gaps due to absent RCT and limited real-world evidence, we applied a doubly-robust semi-supervised estimation approach to registry-linked EHR data to compare the effectiveness of two commonly prescribed DMT classes (BCD v. NTZ) in managing disability progression in pwMS. We found no statistically significant differences between BCD and NTZ in either sustained patient-reported disability outcome (worsening or improvement) up to 3-years after treatment initiation.

The study findings address salient evidence gaps as there has been no RCT comparing BCD and NTZ, both perceived as highly effective DMTs. A few prior observational studies compared the effectiveness of BCD versus NTZ in managing EDSS-based disability progression in pwMS. An Italian retrospective study comparing ocrelizumab (n=124) and NTZ (n=157) in treatment-naive pwMS found no significant differences in time to EDSS worsening within three years after treatment initiation.[43] An Australian retrospective study found no significant difference between ocrelizumab (n=310) and NTZ (n=310) in rates of confirmed EDSS progression, with a median follow-up of 1.5 years.[44] In pwMS who switched from fingolimod, a French study found no difference in the rate of EDSS worsening between patients receiving ocrelizumab/rituximab (n=337) and those receiving NTZ (n=403) over a two-year follow-up.[45] Our study reporting no significant difference between BCD and NTZ in either slowing disability worsening or promoting disability improvement complements the prior studies by evaluating BCD as a mechanistic class, by assessing *sustained* disability change in *patient-reported* disability based on PDDS, and by separately assessing disability worsening and improvement.

Our study has several strengths. First, our study uses a doubly-robust, semi-supervised estimation approach. Previous comparative effective studies of MS DMTs that used an inverse probability of treatment weight required correct specification of the treatment model for consistent estimation of the treatment effect. In contrast, doubly-robust methods require correct specification of either the propensity score model or the outcome model, effectively increasing the likelihood of consistent treatment effect estimation. Second, building on the traditional doubly-robust method, the novel semi-supervised approach in this study leverages patients with and without observed outcome data from registry-linked EHR data by applying outcome imputation models to gain meaningful statistical power in the causal analysis. Third, the EHR data encompassing high-dimensional codified and narrative features provided a more comprehensive source for confounder adjustment to balance patient characteristics than prior studies relying on standard covariates. Fourth, pre-trained knowledge graphs of clinical concepts coupled with the adaptive LASSO method that additionally enables penalized regression for covariate selection (beyond the doubly-robust, semi-supervised property of its estimator) crucially informed the inclusion of clinically relevant features in the causal modeling while reducing feature dimensionality. Finally, the primary study outcome of sustained disability changes, based on 3 post-baseline disability scores (separated by at least 6 months between two consecutive scores) after target treatment initiation, more reliably measures relatively long-term disease status. The causal model using this more rigorous outcome definition is less sensitive to noises from transient data fluctuation, thereby increasing the robustness of the findings and the potential generalizability.

Our study also has limitations. First, while analyzing a sustained disability change outcome captures useful temporal information, this approach may overlook important temporal patterns and short-term fluctuations in disability progression within a 3-year follow-up. To partially address this concern, we conducted sensitivity analyses using a 2-year evaluation window and

found consistent results. Second, while the study cohort was based in a large healthcare system and the semi-supervised estimation approach significantly augmented the statistical power, the current study (with comparable or larger sample size than prior studies comparing BCD and NTZ) might remain underpowered. Future studies involving even larger sample size, other independent cohorts, and more pragmatic rater-assessed disability outcomes (*e.g.,* timed 25-foot walk) will be crucial to validate and further improve the generalizability of the study findings.

Conclusion

In summary, this study contributes to the real-world evidence of the comparative effectiveness of two commonly prescribed highly effective DMT classes (BCD vs NTZ) in managing patient-reported sustained disability progression in pwMS. The doubly-robust, semi-supervised estimation approach that incorporates knowledge graph-derived clinically relevant covariates from EHR has the advantage of achieving more consistent treatment effect estimation, gaining meaningful statistical power, and adjusting relevant confounders more comprehensively than prior studies.

Funding Source

The study is supported by NINDS R01NS098023 (Z Xia). Z Xia is also supported by NINDS R01NS124882.

References

1. Ch P, Pw O, E H, et al. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9). doi:10.1056/NEJMoa044397
2. Hauser SL, Bar-Or A, Weber MS, et al. Association of Higher Ocrelizumab Exposure With Reduced Disability Progression in Multiple Sclerosis. *Neurol Neuroimmunol Neuroinflammation*. 2023;10(2):e200094. doi:10.1212/NXI.0000000000200094
3. Hauser SL, Waubant E, Arnold DL, et al. B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. *N Engl J Med*. 2008;358(7):676-688. doi:10.1056/NEJMoa0706383
4. 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(1):e000304. doi:10.1136/bmjno-2022-000304
5. 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(1):e000304. doi:10.1136/bmjno-2022-000304
6. Ra R, Wh S, Pa C, et al. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9). doi:10.1056/NEJMoa044396
7. Rudick RA, Stuart WH, Calabresi PA, et al. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9):911-923. doi:10.1056/NEJMoa044396
8. Hauser SL, Bar-Or A, Comi G, et al. Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med*. 2017;376(3):221-234. doi:10.1056/NEJMoa1601277
9. Montalban X, Hauser SL, Kappos L, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *N Engl J Med*. 2017;376(3):209-220. doi:10.1056/NEJMoa1606468
10. Hauser SL, Bar-Or A, Cohen JA, et al. Ofatumumab versus Teriflunomide in Multiple Sclerosis. *N Engl J Med*. 2020;383(6):546-557. doi:10.1056/NEJMoa1917246
11. 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 Neurol*. 2022;21(8):693-703. doi:10.1016/S1474-4422(22)00209-5
12. Bar-Or A, Grove RA, Austin DJ, et al. Subcutaneous ofatumumab in patients with relapsing-remitting multiple sclerosis: The MIRROR study. *Neurology*. 2018;90(20):e1805-e1814. doi:10.1212/WNL.0000000000005516
13. Polman CH, O'Connor PW, Havrdova E, et al. A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9):899-910. doi:10.1056/NEJMoa044397
14. Sl H, A BO, Ms W, et al. Association of Higher Ocrelizumab Exposure With Reduced Disability Progression in Multiple Sclerosis. *Neurol Neuroimmunol Neuroinflammation*. 2023;10(2). doi:10.1212/NXI.0000000000200094
15. Hauser SL, Waubant E, Arnold DL, et al. B-cell depletion with rituximab in relapsing-remitting multiple sclerosis. *N Engl J Med*. 2008;358(7):676-688. doi:10.1056/NEJMoa0706383
16. 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(1):e000304. doi:10.1136/bmjno-2022-000304

17. L A, C Z, O F, 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(1). doi:10.1136/bmjno-2022-000304
18. Rudick RA, Stuart WH, Calabresi PA, et al. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9):911-923. doi:10.1056/NEJMoa044396
19. Rudick RA, Stuart WH, Calabresi PA, et al. Natalizumab plus interferon beta-1a for relapsing multiple sclerosis. *N Engl J Med*. 2006;354(9):911-923. doi:10.1056/NEJMoa044396
20. Hauser SL, Bar-Or A, Comi G, et al. Ocrelizumab versus Interferon Beta-1a in Relapsing Multiple Sclerosis. *N Engl J Med*. 2017;376(3):221-234. doi:10.1056/NEJMoa1601277
21. Montalban X, Hauser SL, Kappos L, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. *N Engl J Med*. 2017;376(3):209-220. doi:10.1056/NEJMoa1606468
22. Sl H, A BO, Ja C, et al. Ofatumumab versus Teriflunomide in Multiple Sclerosis. *N Engl J Med*. 2020;383(6). doi:10.1056/NEJMoa1917246
23. A S, T F, J B, 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 Neurol*. 2022;21(8). doi:10.1016/S1474-4422(22)00209-5
24. A BO, Ra G, Dj A, et al. Subcutaneous ofatumumab in patients with relapsing-remitting multiple sclerosis: The MIRROR study. *Neurology*. 2018;90(20). doi:10.1212/WNL.0000000000005516
25. Learmonth YC, Motl RW, Sandroff BM, Pula JH, Cadavid D. Validation of patient determined disease steps (PDDS) scale scores in persons with multiple sclerosis. *BMC Neurol*. 2013;13(1):37. doi:10.1186/1471-2377-13-37
26. Xia Z, Secor E, Chibnik LB, et al. Modeling disease severity in multiple sclerosis using electronic health records. *PloS One*. 2013;8(11):e78927. doi:10.1371/journal.pone.0078927
27. Zhang T, Goodman M, Zhu F, et al. Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. *Neurol Neuroimmunol Neuroinflammation*. 2020;7(6):e864. doi:10.1212/NXI.0000000000000864
28. Ahuja Y, Kim N, Liang L, et al. Leveraging electronic health records data to predict multiple sclerosis disease activity. *Ann Clin Transl Neurol*. 2021;8(4):800-810. doi:10.1002/acn3.51324
29. Liang L, Kim N, Hou J, et al. Temporal Trends of Multiple Sclerosis Disease Activity: Electronic Health Records Indicators. *Mult Scler Relat Disord*. 2022;57:103333. doi:10.1016/j.msard.2021.103333
30. 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(11):e2134627. doi:10.1001/jamanetworkopen.2021.34627
31. Robins JM, Rotnitzky A. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *J Am Stat Assoc*. 1995;90(429):122-129. doi:10.2307/2291135
32. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960. doi:10.1002/sim.1903
33. 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(12):3426-3444. doi:10.1038/s41596-019-0227-6
34. Yu S, Cai T, Cai T. NILE: Fast Natural Language Processing for Electronic Health Records. Published online July 16, 2019. doi:10.48550/arXiv.1311.6063

35. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(Database issue):D267-D270. doi:10.1093/nar/gkh061
36. Xiong X, Sweet SM, Liu M, et al. Knowledge-Driven Online Multimodal Automated Phenotyping System. *MedRxiv Prepr Serv Health Sci*. Published online October 2, 2023:2023.09.29.23296239. doi:10.1101/2023.09.29.23296239
37. Salter A, Lancia S, Cutter G, et al. Characterizing Long-term Disability Progression and Employment in NARCOMS Registry Participants with Multiple Sclerosis Taking Dimethyl Fumarate. *Int J MS Care*. 2021;23(6):239-244. doi:10.7224/1537-2073.2020-109
38. Salter A, Lancia S, Cutter G, et al. A propensity-matched comparison of long-term disability worsening in patients with multiple sclerosis treated with dimethyl fumarate or fingolimod. *Ther Adv Neurol Disord*. 2021;14:17562864211021177. doi:10.1177/17562864211021177
39. Rudick RA, Lee JC, Cutter GR, et al. Disability Progression in a Clinical Trial of Relapsing-Remitting Multiple Sclerosis: Eight-Year Follow-up. *Arch Neurol*. 2010;67(11):1329-1335. doi:10.1001/archneurol.2010.150
40. Cheng D, Chakrabortty A, Ananthakrishnan AN, Cai T. Estimating average treatment effects with a double-index propensity score. *Biometrics*. 2020;76(3):767-777. doi:10.1111/biom.13195
41. Cheng D, Ananthakrishnan AN, Cai T. Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*. 2021;77(2):413-423. doi:10.1111/biom.13298
42. Liu RY, Singh K. Notions of Limiting P Values Based on Data Depth and Bootstrap. *J Am Stat Assoc*. 1997;92(437):266-277. doi:10.1080/01621459.1997.10473624
43. 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. doi:10.1016/j.msard.2024.105594
44. Boz C, Ozakbas S, Terzi M, et al. The comparative effectiveness of fingolimod, natalizumab, and ocrelizumab in relapsing-remitting multiple sclerosis. *Neurol Sci Off J Ital Neurol Soc Ital Soc Clin Neurophysiol*. 2023;44(6):2121-2129. doi:10.1007/s10072-023-06608-z
45. Rollot F, Couturier J, Casey R, et al. Comparative Effectiveness of Natalizumab Versus Anti-CD20 in Highly Active Relapsing-Remitting Multiple Sclerosis After Fingolimod Withdrawal. *Neurother J Am Soc Exp Neurother*. 2022;19(2):476-490. doi:10.1007/s13311-022-01202-1
46. Yuan Z, Zhao Z, Sun H, Li J, Wang F, Yu S. CODER: Knowledge-infused cross-lingual medical term embedding for term normalization. *J Biomed Inform*. 2022;126:103983. doi:10.1016/j.jbi.2021.103983
47. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika*. 2009;96(1):187-199.
48. Althouse AD. Post Hoc Power: Not Empowering, Just Misleading. *J Surg Res*. 2021;259:A3-A6. doi:10.1016/j.jss.2019.10.049