Supplementary Material “The Use and Quality of Reporting of Propensity Score Methods in Multiple Sclerosis Literature: A Review”
Supplementary Methods
Background on Propensity Scores

Propensity Scores
In a study, each subject has observed pre-treatment covariates (L), receives a treatment at baseline, and has an outcome recorded at the end of the follow-up. For convenience, we will refer to A=1 for treated, and A=0 for control; ‘control’ refers to a subject who is untreated or is receiving a placebo or standard therapy. In an observational setting, (a subset of) the covariates under consideration may be confounders; that is, they are informative of both the probability of being treated and of the outcome. As a result, differences in the outcome between the two treatment groups may be attributable to differences in the distribution of these covariates between the two groups rather than to the treatment itself. Imbalance in covariate distributions across treatment arms means that the relationship between treatment assignment and outcome may be confounded, preventing any causal conclusions about the effect of the treatment on the outcome.

The PS is defined as P(A=1|L), the probability of being treated given a subject’s baseline covariates. The PS is a balancing score, meaning that conditioning on the PS balances the distribution of measured covariates between the treatment groups, thus removing the dependency between treatment assignment and outcome \(^1\). Three assumptions are necessary for this property to hold: (i) a subject’s outcome is not influenced by other subjects’ treatment assignments, (ii) no unmeasured confounders, meaning that all covariates that predict both treatment assignment and outcome are recorded in L, and (iii) positivity \(^2\), meaning that each subject has a positive probability of being treated or not at every combination of the confounders.

Estimation of Propensity Scores
In practice, the PS is unknown and must be estimated. Usually, the probability of being treated is modeled using logistic regression with the observed covariates as predictors. An analysis based on the estimated PS assumes that the PS model is correctly specified; that is, sufficient covariates to control confounding are included in the model and in their correct functional
forms. This assumption may be relaxed in some cases by the use of tree-based machine learning methods or other flexible methods. Interaction and high order (polynomial) terms are usually added to the PS model if researchers suspect that a non-linear or non-additive relationship should be accounted in the PS model, which is often evident from plotting relevant bivariate relationship. For a covariate that is not balanced after fitting a simple main effect PS model (e.g., when the respective SMD is still high after PS matching or weighting), changing the specification of the PS model (e.g., adding non-additive or non-linear terms of that covariate) is often helpful to achieve better balance.

Selection of Covariates

Ideally, the selection of the covariates to be included in the PS model should be based on expert opinion and evidence from the literature. The general recommendation is to include covariates that are causally related to the outcome, irrespective of whether they are causally related to the exposure or not, as this increases the likelihood of controlling for confounding while not increasing variance. In the absence of knowledge about causal relationships, several confounder selection criteria are proposed. Recently, automatized variable selection procedures are gaining popularity, in particular, in the context of high-dimensional settings.

Types of Propensity Score Analyses

Four analyses using the PS are commonly used in MS literature: matching, weighting, stratification, and regression adjustment.

Matching

Matching creates pairs (or sets) of treated and control patients who have “similar” values of the PS and compares the outcome within pairs to study the effect of the treatment. Pair matching or 1:1 matching matches one control patient to one treated patient while many-to-one matching, or 1:M matching, finds M control matches for each treated patient. The matched sample refers to the subset of patients successfully matched. Its size may differ from the original sample size, depending on the algorithm used to create matches. The choice of outcome analysis after matching is then driven by the type of outcome (binary, continuous, time-to-event). It is recommended that inferences for the treatment effect should account for
the dependencies within matched sets created via matching by conducting matched pairs analysis (i.e., by calculating robust standard errors, accounting for the clustering in matched pairs) \(^\text{11}\), although such an analysis may lead to loss of power \(^\text{12}\).

Among many more options, two matching algorithms are popular in clinical applications: greedy and optimal matching \(^\text{13}\). In greedy matching, each treated patient is matched to the control patient who has the nearest estimated PS value. A caliper width may be used to define the maximum allowable distance between the estimated PSs of the matched treated and control patients, e.g., 0.05, 0.1, or a fraction (usually 0.2) of the standard deviation (SD) of the PS on the logit scale. A treated patient may remain unmatched and therefore be deleted from the matched sample if no control patient with a similar estimated PS (e.g., within a pre-specified caliper width) can be found. Unlike greedy matching, optimal matching finds the set of matched pairs that minimizes the global distance in estimated PS across all matched pairs.

Greedy and optimal matching can be implemented with or without replacement. Matching with replacement implies that a control patient may be matched to more than one treated patient while matching without replacement uses each control patient as a match at most once. Full matching can be achieved with either algorithm. In full matching, all patients in the original sample are used by matching each control patient to one or more treated patients in addition to matching each treated patient to one or more control patients. The size of the matched sets varies depending on the number of suitable matches available for each patient, and constraints can be imposed to bound the ratio of treated and control subjects in each matched set \(^\text{14}\).

**Weighting**

The estimated PS can be used to derive weights; weighting the sample creates a pseudo-population in which the distribution of the covariates matches some target population. Different choices of weights yield different target populations \(^\text{15}\). The most common weights are the inverse probability of treatment weights in which each patient is weighted by the inverse of the probability of being assigned to their observed treatment, i.e., the PS for treated patients and one minus the PS for control patients. In the weighted sample, the distribution of the
confounders is independent of treatment assignment and matches the distribution of the overall population.

The existence of substantially large weights (e.g., compared to sample size) often leads to increased variability of the estimated effect. Stabilized weights may be computed to reduce the variance of the estimated effect \(^{16,17}\). Alternatively, ad hoc adjustment methods such as weight trimming or truncation are popularly used to reduce the impact of the extreme weights, i.e., weights larger than a fixed threshold are truncated to that threshold \(^{17,18}\).

Stratification
Stratification uses the estimated PS to create strata; for example, one may classify subjects into five strata based on quintiles of the estimated PS distribution \(^{19}\). Stratum-specific estimated treatment effects are subsequently combined into an overall estimated treatment effect.

Regression or Covariate Adjustment
Regression adjustment includes the estimated PS as a covariate in a regression of the outcome on the treatment group indicator. Compared to standard regression that adjusts for each confounder, regression adjustment reduces the dimension of the confounders into a single variable (the estimated PS) at the cost of losing the interpretability of the associated regression coefficient. Unlike the three other approaches, regression adjustment generally assumes that there exists a linear relationship between the outcome and the estimated PSs, and does not directly aid in achieving balance of the covariate distributions (e.g., no clear separation between the design and analysis stages) \(^{20}\). Therefore, the ability of regression adjustment to remove systematic differences between treatment groups is limited compared to other approaches \(^{21}\), although some extensions can improve performance \(^{22,23}\).

Estimands
Different PS methods target different causal treatment effects \(^{20,24}\). The most common causal estimands are the average treatment effect (ATE), which is the treatment effect in the entire population (treated and control), and the average treatment effect in the treated (ATT), which is the effect in the treated group. The choice of an appropriate estimand depends on the research question as well as the target population of interest.
Generally, matching estimates the ATT. Weighting can estimate both the ATE (with inverse probability of treatment weights) or the ATT (by weighting only the control patient by the ratio of the PS over one minus the PS). Stratification can also estimate both ATT and ATE, depending on how stratum-specific estimates are pooled.\(^{25}\)

**Standard Errors**

Calculation of the standard error of the treatment effect estimator should account for (1) the uncertainty of PS estimation (because the true PS is unknown and is estimated with the data), (2) the within-set correlation that may have been induced by matching (for PS matching), and (3) the within-subject correlation in weighted observations (for PS weighting). Ignoring these factors may lead to biased estimation of the standard error, which in turn leads to invalid confidence intervals and inferences. Methods to account for these factors include calculating robust standard errors, which incorporate the uncertainty of the PS estimation. Alternatively, one could use computationally intensive bootstrap-based methods.\(^{17}\)

**Diagnostics**

The balance of the confounders within matched sets, in the weighted sample or across strata, can be assessed with a diagnostic tool popularly known as standardized mean differences (SMDs) for covariates. A successful balance is then inferred if SMDs are below a chosen cut-point, usually 0.1 or 0.2.\(^ {25}\) Graphical display of SMDs (e.g., side-by-side boxplots or empirical distribution functions, or love plot\(^ {21,27}\)) are often used to assess the balance. Formal statistical evaluations such as C-statistics, goodness-of-fit test, Kolmogorov-Smirnov test, or t-test for continuous covariates are sometimes used but many researchers criticize the significance of test-based approaches due to their reliance on sample size.\(^ {24,29}\) Other than checking balance, researchers often report various sensitivity analyses, for example, to estimate the amount of hidden bias due to unmeasured confounding, i.e., the likelihood that an unmeasured confounder would change the conclusion of the analysis.\(^ {30,31}\)

**Additional Methods**

The review was conducted through December 2018 and updated in July 2019 with the most recent articles. The following two terms were used for the database search: “propensity score”
and “multiple sclerosis.” A single reviewer was responsible for the eligibility screening and retrieval of data to a pilot-tested form. The form was pilot-tested, reviewed, and refined after 5, 10 and 25 data extractions to ensure that all necessary data elements were captured and adequately reported. The review team revised the extracted data in August 2019, and one reviewer performed additional extractions in September 2019 to include one missing study. The reviewers were not blinded to the journal and authors. The search period from 2013 to July 2019 was carefully selected based on an assessment of the accessibility and knowledge of PS methods by researchers in MS. For example, a PubMed search using the keywords {multiple sclerosis} and {propensity score found less than 5 articles published per year for the years 2006 to 2012, and more than 5 articles for subsequent years. This suggests that PS methods were still relatively new to the field prior to 2013 and became mainstream starting from 2013. Thus, we have decided that reviewing the articles published from 2013 seems fair with respect to methodological and practical advancements of PS methods in the MS research community.

Supplementary Results
Studies Included in the Review
Thirty-nine of 65 articles were retained for data extraction. Eight articles were published in Multiple Sclerosis Journal, three in Neurology, three in Multiple Sclerosis and Related Disorders, and two or fewer in 19 other peer-reviewed journals. Supplementary Figure 1 gives details on reasons for excluding 26 articles.

General Data Analysis Approaches
Supplementary Table 5 presents the general methodologies adopted in the studies. Only 8% of studies used single or multiple data imputation to address missing data. In 13% of studies, treatment discontinuation and drop-outs were not mentioned. In the remaining 34 studies, treatment persistence was an eligibility criterion (10 studies), was measured or tracked as an outcome of interest (9 studies), was accounted for by censoring the patients at the time of discontinuation or drop-out (22 studies), or was handled via pairwise censoring in studies that used matching (7 of 28 studies). One study mentioned the possibility of treatment
discontinuation but conducted an intention-to-treat analysis. About 64% of the studies reported at least one sensitivity analysis. The influence of unmeasured confounding was tested with sensitivity analyses in 34% of studies, with 21% of those studies using Rosenbaum bounds. About 11% of studies performed on-treatment sensitivity analyses to assess assumptions made about treatment discontinuation. On the extreme end, the presence of the ‘zero-cell’ issues (i.e., no subjects in at least one cell of the confounder categories vs. the treatment group tabulation), a potential (practical) violation of the positivity assumption, was apparent in 13% of studies, whereas the existence of such issue was difficult to determine in 15% of studies. In 54% of studies, the PS adjusted analysis was compared to a basic unadjusted outcome regression. Most studies (85%) explicitly reported the statistical software used for the analysis. R (R Core Team: Vienna, Austria) was the most commonly used software (51%), followed by SPSS (IBM Corp.: Armonk, NY, USA) in 21% of studies, and Stata (StataCorp LP.: College Station, TX, USA) in 18% of studies.

Comparison with Other Reviews
Supplementary Table 6 summarizes the findings from previous reviews on the use and reporting quality of PS in the following disease areas: cardiovascular, infective endocarditis, intensive care and anesthesiology, psychological and educational research, sepsis, critical care, and cancer. Three reviews only focused on studies that used PS matching while two other studies only reviewed studies that could be matched to an RCT.

Matching was the most commonly used method across studies reviewed in all disease areas. Our review found the highest proportion of studies that used weighting as the PS method (15%) compared to the other reviews, which were more likely to report adjustment or stratification as PS method. Our comparison with other reviews also highlighted good practices and opportunities for improvement. For studies using matching, between 19% and 35% of the studies in previous reviews did not report the matching algorithm that was used. This proportion was only 7% in our MS review. Our review reported the smallest median number of variables used in the PS model despite reporting relatively large sample sizes and numbers of treated patients. Further, there was a relatively high sample size reduction due to matching in our review (average: 46%) and previous reviews (32-56%). Encouragingly, a high proportion of
PS matching studies checked the post-PS balance in our review (96%) as well as previous reviews (50% to 88%). Future reviews of PS methods should extract key elements of the overall analysis approach to allow better comparisons between reviews. These reviews should include sample size, balance diagnostics, sample size reduction, treated-to-variable ratio (i.e., number of treated subjects divided by the number of covariates included in the model), and matching algorithm.

**Supplementary Discussion**

*Areas of Potential Knowledge Gaps*

From the PS literature, it is known that even slight PS model misspecification may result in a biased treatment effect estimate \(^78\). In real-world applications, analysts are generally trying to guess the correct specification of the PS model, while trying to obtain covariate balance. While estimating the PS, the prediction can be improved by including a non-linear function of the confounders or with interaction terms in the logistic regression. Robustness to PS model misspecification can also be improved by using machine learning and ensemble methods, in terms of PS prediction, the bias in the effect estimate, and covariate balance \(^79\). These alternative methods for fitting PS models are not commonly practiced within the MS research community.

Our review also highlighted that a large proportion of studies used p-values to check for post-PS imbalances (43% of studies that used matching and 35% of studies that used weighting), which is generally discouraged.
Supplementary References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41-55.

2. Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research.* 2012;21(1):31-54.

3. Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology.* 2010;63(8):826-833.

4. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine.* 2013;32(19):3388-3414.

5. Kainz K, Greifer N, Givens A, et al. Improving Causal Inference: Recommendations for Covariate Selection and Balance in Propensity Score Methods. *Journal of the Society for Social Work & Research.* 2017;8(2):279-303.

6. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *American Journal of Epidemiology.* 2006;163(12):1149-1156.

7. VanderWeele TJ. Principles of confounder selection. *European Journal of Epidemiology.* 2019;34(3):211-219.

8. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512-522.

9. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine.* 2010;29(3):337-346.

10. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiology.* 2018;29(2):191-198.

11. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *The Journal of Thoracic and Cardiovascular Surgery.* 2007;134(5):1128-1135.

12. Hill J. Discussion of research using propensity-score matching: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’ by Peter Austin, Statistics in Medicine. *Statistics in Medicine.* 2008;27(12):2055-2061.

13. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine.* 2014;33(6):1057-1069.

14. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association.* 2004;99(467):609-618.

15. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association.* 2018;113(521):390-400.

16. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology.* 2008;168(6):656-664.
17. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine.* 2015;34(28):3661-3679.
18. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PloS One.* 2011;6(3):e18174.
19. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association.* 1984;79(387):516-524.
20. Ali MS, Groenwold RH, Klungel OH. Best (but oft-forgotten) practices: propensity score methods in clinical nutrition research. *American Journal of Clinical Nutrition.* 2016;104(2):247-258.
21. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making.* 2009;29(6):661-677.
22. Tian Y, Baro E, Zhang R. Performance evaluation of regression splines for propensity score adjustment in post-market safety analysis with multiple treatments. *Journal of Biopharmaceutical Statistics.* 2019;29(5):810-821.
23. Franklin JM, Eddings W, Austin PC, Stuart EA, Schneeweiss S. Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Statistics in Medicine.* 2017;36(12):1946-1963.
24. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research.* 2011;46(1):119-151.
25. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science: a review journal of the Institute of Mathematical Statistics.* 2010;25(1):1-21.
26. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine.* 2008;27(12):2037-2049.
27. Rudolph KE, Stuart EA, Glass TA, Merikangas KR. Neighborhood disadvantage in context: The influence of urbanicity on the association between neighborhood disadvantage and adolescent emotional disorders. *Social Psychiatry and Psychiatric Epidemiology.* 2014;49(3):467-475.
28. Ali MS, Groenwold RH, Belitser SV, et al. Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review. *Journal of Clinical Epidemiology.* 2015;68(2):122-131.
29. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A.* 2008;171(2):481-502.
30. Rosenbaum PR. Sensitivity to hidden bias. In: *Observational studies.* Springer, New York; 2002:105-170.
31. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological).* 1983;45(2):212-218.
Alsop J, Medin J, Cornelissen C, Vormfelde SV, Ziemssen T. Two studies in one: A propensity-score-matched comparison of fingolimod versus interferons and glatiramer acetate using real-world data from the independent German studies, PANGAEA and PEARL. *PloS One.* 2017;12(5):e0173353.

Barbin L, Rousseau C, Jousset N, et al. Comparative efficacy of fingolimod vs natalizumab: A French multicenter observational study. *Neurology.* 2016;86(8):771-778.

Baroncini D, Ghezzi A, Annovazzi PO, et al. Natalizumab versus fingolimod in patients with relapsing-remitting multiple sclerosis non-responding to first-line injectable therapies. *Multiple Sclerosis Journal.* 2016;22(10):1315-1326.

Bergvall N, Lahoz R, Reynolds T, Korn JR. Healthcare resource use and relapses with fingolimod versus natalizumab for treating multiple sclerosis: A retrospective US claims database analysis. *Current Medical Research and Opinion.* 2014;30(8):1461-1471.

Braune S, Lang M, Bergmann A. Efficacy of fingolimod is superior to injectable disease modifying therapies in second-line therapy of relapsing remitting multiple sclerosis. *Journal of Neurology.* 2016;263(2):327-333.

Chalmer TA, Kalincik T, Laursen B, Sorensen PS, Magyari M. Treatment escalation leads to fewer relapses compared with switching to another moderately effective therapy. *Journal of Neurology.* 2019;266(2):306-315.

Chen C, Wu N, Watson C. Multiple sclerosis patients who are stable on interferon therapy show better outcomes when staying on same therapy than patients who switch to another interferon. *ClinicoEconomics and Outcomes Research.* 2018;10:723-730.

Cocco E, Sardu C, Spinicci G, et al. Influence of treatments in multiple sclerosis disability: A cohort study. *Multiple Sclerosis Journal.* 2015;21(4):433-441.

Cohan SL, Moses H, Calkwood J, et al. Clinical outcomes in patients with relapsing-remitting multiple sclerosis who switch from natalizumab to delayed-release dimethyl fumarate: A multicenter retrospective observational study (STRATEGY). *Multiple Sclerosis and Related Disorders.* 2018;22:27-34.

D’Amico E, Zanghì A, Callari G, et al. Comparable efficacy and safety of dimethyl fumarate and teriflunomide treatment in relapsing-remitting multiple sclerosis: An Italian real-word multicenter experience. *Therapeutic Advances in Neurological Disorders.* 2018;11:1756286418796404.

D’Amico E, Patti F, Zanghì A, Lo Fermo S, Chisari C, Zappia M. Lateral switch to IFN beta-1a 44 mcg may be effective as escalation switch to fingolimod in selected persons with relapsing remitting multiple sclerosis: A real-world setting experience. *Expert Review of Clinical Pharmacology.* 2018;11(5):531-536.

Guger M, Enzinger C, Leutmezer F, et al. Real-life clinical use of natalizumab and fingolimod in Austria. *Acta Neurologica Scandinavica.* 2018;137(2):181-187.

He A, Spelman T, Jokubaitis V, et al. Comparison of switch to fingolimod or interferon beta/glatiramer acetate in active multiple sclerosis. *JAMA Neurology.* 2015;72(4):405-413.

Healy BC, Glanz BI, Zurawski JD, Mazzola M, Chitnis T, Weiner HL. Long-term follow-up for multiple sclerosis patients initially treated with interferon-beta and glatiramer acetate. *Journal of the Neurological Sciences.* 2018;394:127-131.
46. Hersh CM, Love TE, Bandyopadhyay A, et al. Comparative efficacy and discontinuation of dimethyl fumarate and fingolimod in clinical practice at 24-month follow-up. *Multiple Sclerosis Journal–Experimental, Translational and Clinical.* 2017;3(3):2055217317715485.

47. Iaffaldano P, Lucisano G, Pozzilli C, et al. Fingolimod versus interferon beta/glatiramer acetate after natalizumab suspension in multiple sclerosis. *Brain.* 2015;138(11):3275-3286.

48. Johnson BH, Bonafede MM, Watson C. Platform therapy compared with natalizumab for multiple sclerosis: Relapse rates and time to relapse among propensity score-matched US patients. *CNS Drugs.* 2015;29(6):503-510.

49. Kalincik T, Spelman T, Trojano M, et al. Persistence on therapy and propensity matched outcome comparison of two subcutaneous interferon beta 1a dosages for multiple sclerosis. *PLoS One.* 2013;8(5):e63480.

50. Kalincik T, Horakova D, Spelman T, et al. Switch to natalizumab versus fingolimod in active relapsing–remitting multiple sclerosis. *Annals of Neurology.* 2015;77(3):425-435.

51. Kalincik T, Jokubaitis V, Spelman T, et al. Cladribine versus fingolimod, natalizumab and interferon β for multiple sclerosis. *Multiple Sclerosis Journal.* 2018;24(12):1617-1626.

52. Kalincik T, Brown JWL, Robertson N, et al. Treatment effectiveness of alemtuzumab compared with natalizumab, fingolimod, and interferon beta in relapsing-remitting multiple sclerosis: A cohort study. *The Lancet Neurology.* 2017;16(4):271-281.

53. Kister I, Spelman T, Alroughani R, et al. Discontinuing disease-modifying therapy in MS after a prolonged relapse-free period: A propensity score-matched study. *Journal of Neurology, Neurosurgery and Psychiatry.* 2016;87(10):1133-1137.

54. Koch-Henriksen N, Magyari M, Sellebjerg F, Soelberg Sørensen P. A comparison of multiple sclerosis clinical disease activity between patients treated with natalizumab and fingolimod. *Multiple Sclerosis Journal.* 2017;23(2):234-241.

55. Lanzillo R, Carotenuto A, Moccia M, et al. A longitudinal real-life comparison study of natalizumab and fingolimod. *Acta Neurologica Scandinavica.* 2017;136(3):217-222.

56. Lorscheider J, Kuhle J, Izquierdo G, et al. Anti-inflammatory disease-modifying treatment and disability progression in primary progressive multiple sclerosis: A cohort study. *European Journal of Neurology.* 2019;26(2):363-370.

57. Lorscheider J, Jokubaitis VG, Spelman T, et al. Anti-inflammatory disease-modifying treatment and short-term disability progression in SPMS. *Neurology.* 2017;89(10):1050-1059.

58. Lorscheider J, Benkert P, Lienert C, et al. Comparative analysis of natalizumab versus fingolimod as second-line treatment in relapsing–remitting multiple sclerosis. *Multiple Sclerosis Journal.* 2018;24(6):777-785.

59. Moccia M, Palladino R, Carotenuto A, et al. A 8-year retrospective cohort study comparing Interferon-β formulations for relapsing-remitting multiple sclerosis. *Multiple Sclerosis and Related Disorders.* 2018;19:50-54.

60. Ontaneda D, Nicholas J, Carraro M, et al. Comparative effectiveness of dimethyl fumarate versus fingolimod and teriflunomide among MS patients switching from first-generation platform therapies in the US. *Multiple Sclerosis and Related Disorders.* 2019;27:101-111.
61. Prosperini L, Lucchini M, Haggiag S, et al. Fingolimod vs dimethyl fumarate in multiple sclerosis: A real-world propensity score-matched study. *Neurology.* 2018;91(2):e153-e161.

62. Spelman T, Kalincik T, Zhang A, et al. Comparative efficacy of switching to natalizumab in active multiple sclerosis. *Annals of Clinical and Translational Neurology.* 2015;2(4):373-387.

63. Spelman T, Kalincik T, Jokubaitis V, et al. Comparative efficacy of first-line natalizumab vs IFN-β or glatiramer acetate in relapsing MS. *Neurology: Clinical Practice.* 2016;6(2):102-115.

64. Spelman T, Frisell T, Piehl F, Hillert J. Comparative effectiveness of rituximab relative to IFN-β or glatiramer acetate in relapsing-remitting MS from the Swedish MS registry. *Multiple Sclerosis Journal.* 2018;24(8):1087-1095.

65. Vollmer B, Nair KV, Sillau SH, Corboy J, Vollmer T, Alvarez E. Comparison of fingolimod and dimethyl fumarate in the treatment of multiple sclerosis: Two-year experience. *Multiple Sclerosis Journal–Experimental, Translational and Clinical.* 2017;3(3):2055217317725102.

66. Vollmer B, Ontaneda D, Bandypadhyay A, et al. Discontinuation and comparative effectiveness of dimethyl fumarate and fingolimod in 2 centers. *Neurology: Clinical Practice.* 2018;8(4):292-301.

67. Zecca C, Merlini A, Disanto G, et al. Half-dose fingolimod for treating relapsing-remitting multiple sclerosis: Observational study. *Multiple Sclerosis Journal.* 2018;24(2):167-174.

68. Zhang T, Shirani A, Zhao Y, et al. Beta-interferon exposure and onset of secondary progressive multiple sclerosis. *European Journal of Neurology.* 2015;22(6):990-1000.

69. Péron J, Buyse M, Ozenne B, Roche L, Roy P. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Statistical Methods in Medical Research.* 2018;27(4):1230-1239.

70. Rosenbaum PR. Overt bias in observational studies. In: *Observational studies.* Springer, New York; 2002:71-104.

71. McMurry TL, Hu Y, Blackstone EH, Kozower BD. Propensity scores: methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. *The Journal of Thoracic and Cardiovascular Surgery.* 2015;150(1):14-19.

72. Tleyjeh IM, Kashour T, Zimmerman V, Steckelberg JM, Wilson WR, Baddour LM. The role of valve surgery in infective endocarditis management: a systematic review of observational studies that included propensity score analysis. *American Heart Journal.* 2008;156(5):901-909.

73. Gayat E, Pirracchio R, Resce-Rigon M, Mebazaa A, Mary J-Y, Porcher R. Propensity scores in intensive care and anaesthesiology literature: A systematic review. *Intensive Care Medicine.* 2010;36(12):1993-2003.

74. Thoemmes FJ, Kim ES. A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research.* 2011;46(1):90-118.

75. Zhang Z, Ni H, Xu X. Do the observational studies using propensity score analysis agree with randomized controlled trials in the area of sepsis? *Journal of Critical Care.* 2014;29(5):886.e889-886.e815.
76. Zhang Z, Ni H, Xu X. Observational studies using propensity score analysis underestimated the effect sizes in critical care medicine. *Journal of Clinical Epidemiology*. 2014;67(8):932-939.

77. Yao XI, Wang X, Speicher PJ, et al. Reporting and guidelines in propensity score analysis: A systematic review of cancer and cancer surgical studies. *Journal of the National Cancer Institute*. 2017;109(8):djw323.

78. Kang JD, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*. 2007;22(4):523-539.

79. Pirracchio R, Petersen ML, van der Laan M. Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*. 2014;181(2):108-119.
Supplementary Figures

Supplementary Figure 1. Flowchart describing reasons for excluding articles identified based on title and abstract screening.
**Supplementary Table 1.** General statistical considerations in the 39 studies reviewed

| Characteristics                                      | No. studies/total (%) |
|------------------------------------------------------|-----------------------|
| **Missing data imputation**                          |                       |
| Single                                               | 2/39 (5)              |
| Multiple                                             | 1/39 (3)              |
| Not reported                                         | 36/39 (92)            |
| **How was treatment discontinuation integrated?**    |                       |
| Eligibility criteria                                 | 10/39 (26)            |
| Outcome                                              | 9/39 (23)             |
| Censoring                                            | 22/39 (56)            |
| Not mentioned                                        | 5/39 (13)             |
| **Sensitivity analyses**                             |                       |
| Yes                                                  | 25/39 (64)            |
| No                                                   | 14/39 (36)            |
| **Robust standard error**                            |                       |
| Yes                                                  | 9/39 (23)             |
| Not reported                                         | 30/39 (77)            |
| **Software^a**                                       |                       |
| R                                                    | 20/39 (51)            |
| SPSS                                                 | 8/39 (21)             |
| Stata                                                | 7/39 (18)             |
Eight studies used more than one software.

SAS Institute: Cary, NC, USA.

Statsoft: Tulsa, OK, USA.

PS: propensity score, SMD: standardized mean difference.

Supplementary Table 2. Comparison with propensity score analyses reviews in other disease areas

| Disease area               | Austin 2007 11  | Tleyjeh et al. 2008 72 | Gayat et al. 2010 73 | Thoemmes et al. 2011 74 | Zhang et al. 2014 75 | Zhang et al. 2014 76 | McMurry et al. 2015 71 | Yao et al. 2017 77 | Present study |
|----------------------------|-----------------|------------------------|----------------------|-------------------------|----------------------|----------------------|----------------------|-------------------|---------------|
| Period                     | 2004-2006       | 2003-2007              | 2006-2009            | 2003-2009               | 2007-2013            | 2005-2013            | 2013-2014            | 2002-2015       | 2013-2019     |
| No. articles reviewed      | 60              | 6                      | 47                   | 86                      | 14                   | 20                   | 89                   | 339              | 39            |
| Median No. patients (Q1-Q3)| NA              | 434 (355-546)          | 2,186 (498-5,612)    | NA                      | 829 (178-6,027)      | 433 (181-878)        | NA                   | CS: 4,515 (1,392-20,600) | 951 (563-2,557) |
| Median No. variables in PS model (Q1-Q3) | NA | Min.: 5 | 15 (9-22) | Reported in 37 studies: 15 (10-29) | 11 (8-24) | NA | Reported in 308 studies | 7 (6-9) |
| Treated-to-variable ratio < 10 in PS | NA | NA | 11 (31) | NA | NA | NA | NA | NA | 3 (8) |
|-------------------------------------|----|----|----------|----|----|----|----|----|------|
| **PS method**                       |    |    |          |    |    |    |    |    |      |
| **Matching**                        | 60 (100) | 6 (100) | 26 (55) | 58 (67) | 8 (57) | 14 (70) | 89 (100) | 240 (71) | 28 (72) |
| **Weighting**                       | 0 | 0 | 0 | 6 (7) | 0 | 1 (5) | 0 | 29 (9) | 6 (15) |
| **Adjustment**                      | 0 | 0 | 9 (19) | 4 (5) | 6 (43) | 5 (25) | 0 | 32 (9) | 7 (18) |
| **Stratification**                  | 0 | 0 | 12 (26) | 21 (24) | 1 (7) | 0 | 0 | 28 (8) | 1 (3) |
| **If matching**                     |    |    |          |    |    |    |    |    |      |
| **Balance checked**                 | 49 (82) | 3 (50) | 20 (77) | NA | 7 (88) | NA | 66 (74) | 211 (88) | 27 (96) |
| **With p-values**                   | 47 (96) | NA | 14 (70) | NA | NA | NA | NA | 56 (23) | 12 (44) |
| **Matching algorithm**              |    |    |          |    |    |    |    |    |      |
| **Greedy**                          | 43 (72) | NA | 21 (81) | 31 (53) | NA | NA | NA | 148 (62) | 26 (93) |
| **Not reported**                    | 17 (28) | NA | 5 (19) | 13 (22) | NA | NA | 29 (33) | 83 (35) | 2 (7) |
| **Median sample size reduction (Q1,Q3) (%)** | NA | 32 (24-54) | NA | 56 (47-74) | NA | NA | NA | 42 (33-62) |

Results reported as frequency (%) or median (Q1–Q3) unless stated otherwise.

CS: cancer studies, CSS: cancer surgery studies, Max.: maximum, Min.: minimum, No.: number, PS: propensity score, NA: Not available.