

Observational Studies: Propensity Score Analysis of Non-randomized Data

M Trojano¹, F Pellegrini², D Paolicelli¹, A Fuiani¹, V Di Renzo¹

¹Department of Neurological and Psychiatric Sciences, University of Bari, Bari, Italy; ²Department of Clinical Pharmacology and Epidemiology, Consorzio Mario Negri Sud, Santa Maria Imbaro, Chieti, Italy

Summary

The randomized controlled trial (RCT) is considered to be the 'gold standard' for providing evidence on drug efficacy. However, particularly for answering long-term questions in chronic diseases such as multiple sclerosis (MS), RCTs are often not feasible because of their size, duration, ethical constraints and costs. Data derived from observational studies complement information provided by RCTs. A major issue is that observational studies are more exposed and prone to biases, which can partly be addressed through rigorous study design or statistical

analysis. Propensity score (PS) techniques are the most frequently used. PS is the probability that an individual would receive a certain treatment based on his/her pre-treatment characteristics. This score is being widely used in many therapeutic areas and also in MS to adjust for the uncontrolled assignment of treatment in observational studies. However, since PS cannot adjust for unmeasured or unknown confounders, the conclusions from an observational study may not be considered as strong as those from RCTs.

KEY WORDS:

PROPENSITY SCORE; OBSERVATIONAL STUDY; RANDOMIZED CONTROLLED TRIAL; MULTIPLE SCLEROSIS

Introduction

Randomized controlled trials (RCTs) provide the strongest evidence for the efficacy of preventive and therapeutic procedures in the clinical setting.¹ In RCTs, subjects are randomized to a treatment or control group. By the very nature of chance, random assignment tends to balance covariates, so that there are not systematic differences (bias) in measured and unmeasured covariates between subjects assigned to treated and control groups. If randomization is performed correctly, differing outcomes indicate treatment effect.² However, RCTs do not necessarily provide the definitive answer to drug effectiveness (benefit under the conditions of usual practice or long-term results), as there are many restrictions that limit generalizability³ and there are instances in which RCTs are unethical or impractical. RCTs tend to gather accurate, detailed and standardized information on highly selected subgroups of patients,

treated with a particular drug, at a fixed dosage, and with the greatest potential for benefit over a short time (Table 1). Therefore, after the demonstration of treatment effect under study conditions (efficacy), its applicability to real life (effectiveness) needs to be tested.⁴

The observational study is currently considered a research tool to complement information provided by RCTs.⁵ Moreover, on the one hand, observational studies allow the assessment of the effectiveness of drugs, already proven effective in short-term RCTs (Table 1). In such studies, the typically large patient populations have been exposed to the treatment of interest, often as part of their medical care, and they could potentially benefit over the long term. On the other hand, effectiveness data derived from observational studies may be used as hypothesis generating to be tested later in a RCT or carefully

Table 1: Comparison between RCTs and observational studies

RCTs	Observational studies
Test efficacy in highly selected subgroups with the greatest potential for benefit	Test effectiveness in typical patient population who might potentially benefit
Short-term follow-up	Long-term follow-up
Dosage regimen is often inflexible	Dosing is flexible
One treatment is often used for each patient enrolled	Different treatments can be assigned to the same patient
Small changes of outcomes or surrogate outcomes are often used	Outcomes are usually more robust and can include rare events
Extremely costly	Less expensive
Ethical constraints	

conducted longitudinal studies.⁶ Further, observational studies are useful for establishing long-term safety and detecting rare side-effects of drugs and drug–drug interactions. Moreover, RCTs are impractical for identifying rare events because of the very large number of cases that may be required to find a statistically significant difference between the study arms. Finally, observational studies can provide information when there are substantial barriers to the conduct of RCTs, such as the requirement for an extremely large sample size or a very long period of follow-up – e.g. assessment of treatments for multiple sclerosis (MS). However, economic, regulatory or political obstacles are not sufficient to justify the use of observational studies for providing evidence on drug efficacy. Instead, efforts need to be made to overcome any obstacles that inappropriately prevent the provision of reliable evidence from RCTs of adequate size.^{5,7,8}

Sources of Bias in Observational Studies

A major issue is that observational studies are more exposed and prone to biases than RCTs (Table 2).^{9–11} The fundamental criticism of observational studies, attempting to estimate the effect of a treatment by comparing outcomes for non-randomized subjects, is that either known or unknown confounding factors may influence the measured association between an exposure of interest and a given outcome. Differences in outcomes can be due to differences between the patient groups, in ascertainment of

outcomes, unintended differences in other treatment factors, or to the treatment factor being studied and even the outcome measures.¹² Treatment-by-indication bias may arise when patient characteristics influence drug prescription and, at the same time, also relate to outcome (survival), therefore acting as confounders.¹³

The reasons why certain patients received a treatment while others did not are often difficult to fully account for, and, if these characteristics also affect the outcome, direct comparison of the groups is likely to produce biased conclusions. Further bias may be due to differential detection of outcomes.⁸ Patients receiving any treatment will tend to be seen by doctors or other health professionals more frequently than untreated patients, which may result in the earlier detection of a variety of outcomes.¹⁴

Table 2: Bias in observational studies of treatment

Confounding: Systematic error due to the failure to account for the effect of one or more variables that are related to both the causal factor being studied and the outcome and are not distributed the same between the groups being studied. Confounding occurs when a factor is associated with the use (confounding by indication) or avoidance (confounding by contraindication) of the treatment, but independently influences the risk of the outcome of interest

Recall bias: Systematic error that occurs when the reliability of recall of treatment exposure differs between those who develop an adverse outcome and those who do not

Detection bias: Systematic error that occurs when, because of the lack of blinding or related reasons, the measurement methods are consistently different between groups in the study

Finally, recall bias can be a problem in observational studies when there is a difference in the reliability of the data collected on treatment exposure between cases that have the disease of interest and controls that do not.^{8,15}

A recent paper¹⁶ underlined that important issues related to confounding are often not clearly addressed, from the reader's perspective, in published observational studies. Failure to recognize the limitations of observational studies in the assessment of treatment effects may have serious consequences, including both the use of ineffective or dangerous treatments and the inappropriate abandonment, or insufficiently widespread use, of effective treatments.^{8,17-19}

Propensity Score Analyses to Ensure Validity in the Absence of Randomization

Several possible methodological improvements (regression adjustment, stratification and matching) have been proposed and are available to deal with confounding and to improve validity when randomization is absent.²⁰ Regression analyses estimate the association of each independent variable (baseline characteristics and the intervention) with the dependent variable (outcome of interest) after adjusting for the effects of all the other variables, so that they provide an adjusted estimate of the intervention effect.²⁰ Stratification consists of grouping subjects into strata determined by observed background characteristics believed to confound the analysis. Treated and control subjects in the same strata are compared directly. Stratification creates subgroups that are more balanced in terms of confounders than the total population which can result in less biased estimates of the intervention effect.²⁰ Matching techniques allow to match individual cases (i.e. treated patients) with individual controls that have similar confounding factors in order to reduce the effect of these on the association being investigated in analytical studies. This is most commonly seen in case-control studies and when there are only limited numbers of treated patients and a much larger number of untreated (or control) patients. However, these traditional methods of adjustment are often limited since they can only use a

Table 3: Propensity score

Device for balancing numerous observed covariates^{21,22}

Definition

Formal: The conditional probability of exposure to a treatment given observed covariates

Intuitive: The likelihood that a person would have been treated using only their covariate scores

Collection of covariates is collapsed into a single variable; the probability (or propensity) of being treated.

Limitation

Does not control for unobserved variables that may affect whether subjects receive treatment

small number of covariates for adjustment or if there is extreme imbalance in the background characteristics.^{2,3}

The propensity score (PS) technique, introduced by Rosenbaum and Rubin in 1983,^{21,22} providing a scalar summary measure of the covariate information, lessens this limitation. It represents an alternative method for estimating treatment effects when treatment assignment is not random, but can be assumed to be unconfounded (Table 3). It may be considered as a balancing score that can be used to reduce bias through the adjustment methods mentioned above. Intuitively, the PS is a measure of the probability (between 0 and 1) that an individual is in the 'treated' group given his or her background (pre-treatment) characteristics (e.g. age, disease severity).^{3,23}

The following steps have been proposed to derive a PS model:²⁴

- First, choose a list of background baseline variables that, based on previous available empirical evidence, are likely related to exposure and/or the outcome. For example, in studies^{25,26} aimed to evaluate the impact of a treatment on disability endpoints in MS cohorts, we choose the following variables at treatment assignment: age at disease onset, gender, disease duration, number of relapses in the past year and Expanded Disability Status Scale (EDSS) score.
- Secondly, derive an initial propensity score model by using logistic regression, in which the treatment

variable ('treated', yes or no) is the outcome and the background characteristics (i.e. in MS studies^{25,26} mentioned above, age at disease onset, gender, disease duration, number of relapses in the past year and EDSS score were considered) are the predictor variables in the model. Propensity scores are then calculated for each participant by applying their background values to the logistic model. Propensity scores range from 0 to 1 and reflect each participant's conditional probability of being exposed to treatment given baseline characteristics. Scores close to 1 indicate participant characteristics associated with a high probability of being treated or drug exposed.²⁴

- Thirdly, assess whether matching on the propensity score results in a matched sample in which measured baseline variables are balanced between treated and untreated subjects. Non-overlapping subjects were excluded from the subsequent analyses.

Propensity score methods are increasingly being used in observational studies in which: 1) baseline characteristics differ between the exposed and unexposed groups; 2) exposure is relatively common; 3) the number of measured characteristics or potential confounders is relatively large; and 4) the number of events is relatively small. The most common methods in the medical and epidemiological literature are stratification on the PS,^{27,28} PS matching,²⁷ and covariate adjustment using the PS.^{29,30} With all three techniques, the PS is calculated in the same way,³¹ but once estimated is applied differently. In some of these methods, the PS is used in the analyses as a weight or factor (regression adjustment), whereas in others it is used to construct the appropriate comparisons (stratification or matching), but not in the analyses directly.³

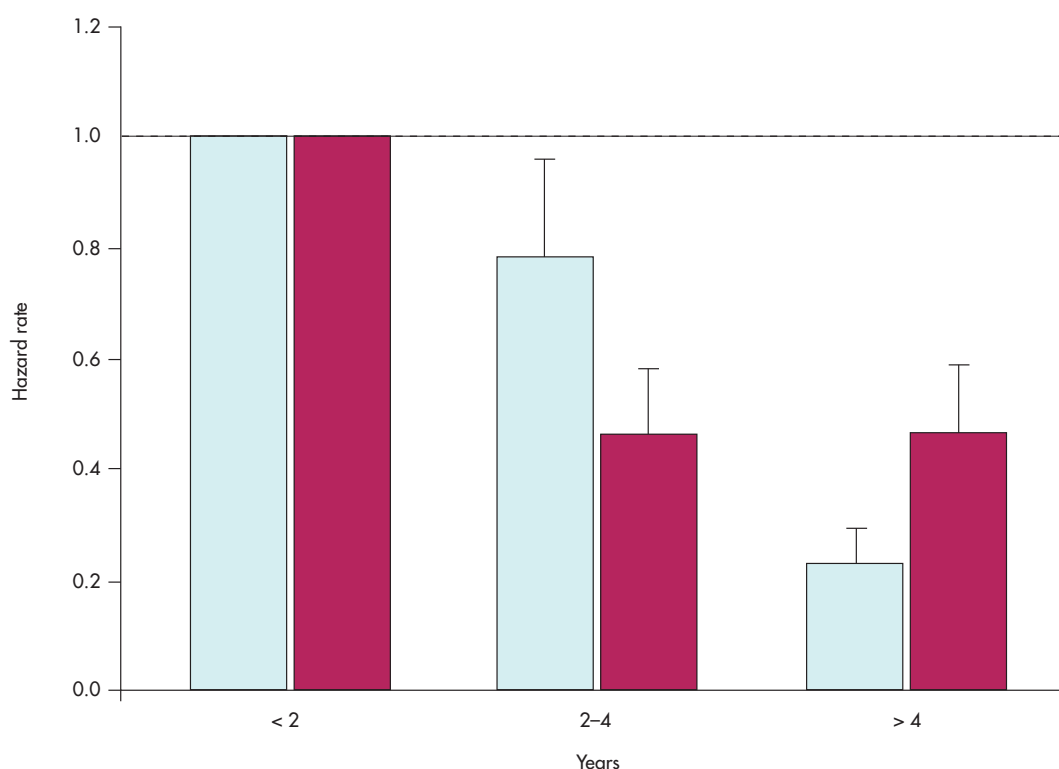
A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects has been performed.³² A more recent paper,³³ by the same authors, examined the performance of different PS methods for estimating relative risks. The authors found that covariate adjustment using the PS tended to have the best performance for estimating relative risks. Matching and stratification on the PS resulted in

Key Points

- *Random assignment balances known and unknown baseline characteristics (covariates) between treatment and control groups*
- *Propensity score represents an alternative method for estimating treatment effects when treatment assignment is not random*
- *The use of the PS can create a 'quasi-randomized' experiment*
- *Hidden bias is the great problem with all methods of adjusting for overt biases, including PS, in observational studies*
- *A sensitivity analysis determine the magnitude of a potential unmeasured confounder that would need to be present to materially alter the conclusions of a study*

estimates with similar mean squared error, with matching resulting in less bias and stratification resulting in estimates with greater precision. Propensity scores are being widely used in pharmacoepidemiological and health-economic analyses, particularly to test drug effects in many therapeutic areas^{34–39} including MS.^{25,26,40}

A Cox proportional hazards regression adjusted for PS was used²⁵ to assess the risk of disability progression and worsening of relapse rate according to the length of exposure to interferon beta (IFNB) in a large cohort of 2090 MS patients collected by the Italian MS Database Network. Forty-one per cent of patients were exposed to IFNB for up to 2 years, 39% for 2–4 years and 20% for more than 4 years. The risks of disability progression (hazard ratio [HR]=0.23; 95% confidence interval [CI]: 0.17–0.30) and worsening of relapse rate (HR=0.19; 95% CI: 0.14–0.27) were reduced by about 4- to 5-fold in patients exposed to IFNB for more than 4 years, compared with patients exposed for up to 2 years. The propensity score technique confirmed these findings (Figure 1). The long-term impact of IFNB treatment on three different clinical end-points: times from first visit to reach an irreversible clinical disability corresponding to Kurtzke's EDSS scores 4 and 6 and to reach secondary progression (SP), was evaluated in a large cohort of untreated (n=401) and IFNB-treated



Blue bars represent hazard rates not adjusted by propensity score; red bars represent adjusted rates. Error bars represent 95% CIs. Adapted from Trojano *et al.*²⁵

Figure 1. Cox regression analysis of disability progression according to level of IFNB exposure and propensity score technique

($n=1103$) relapsing-remitting MS (RRMS) patients using an inverse-weighting PS-adjusted Cox proportional hazards regression.²⁶ Times from first visit and from date of birth were used as survival time variables. The IFNB-treated group showed a highly significant reduction in the incidence of SP (HR=0.38; 95% CI: 0.24–0.58 for time from first visit; HR=0.36; 95% CI: 0.23–0.56 for time from date of birth; $P<0.0001$), EDSS score of 4.0 (HR=0.70; 95% CI: 0.53–0.94 for time from first visit; HR=0.69; 95% CI: 0.52–0.93 for time from date of birth; $P<0.02$), and EDSS score of 6.0 (HR=0.60; 95% CI: 0.38–0.95 for time from first visit; HR=0.54, 95% CI: 0.34–0.86 for time from date of birth; $P\leq 0.03$) when compared with untreated patients (Table 4). SP, EDSS score of 4.0 and EDSS score of 6.0 were reached with significant delays estimated by time from first visit (3.8, 1.7 and 2.2 years, respectively) and time from date of birth (8.7, 4.6 and 11.7 years, respectively) in favour of treated patients. The percentage of patients in the untreated group who converted to SP (20.2%) up to

7 years of follow-up (about 11 years from onset) was in accordance with the estimated mean rate of conversion to SP of 2–3% per year evaluated in previous natural-history studies,⁴¹ whereas the proportions of treated patients who converted to SP was significantly lower (8%). The proportions of untreated patients who reached EDSS scores of 4.0 (27.8%) and 6.0 (12.4%) are in accordance with more recent natural-history data,⁴² whereas the same proportions were lower in treated patients (20.5% and 7.7% for EDSS score of 4.0 and 6.0, respectively).

A PS matching analysis was performed⁴⁰ to assess the robustness of multivariate models in a post-marketing evaluation of the impact of neutralizing antibodies (NAbs) on the effectiveness of IFNB in RRMS. PS-matched Poisson Regression analysis, adjusted at the time in which comparator group first became positive, showed a significant increase of incidence of relapses and a trend to a higher risk of confirmed EDSS 4.0 during the NAb+ in comparison with NAb– status.

In a cohort of 2570 IFNB-treated RRMS patients, a

Table 4: Results of the analysis of time from first visit to the three clinical end-points using propensity score-adjusted Cox models for the IFNB-treated group versus the untreated control group

Survival time	SP			EDSS 4.0			EDSS 6.0		
	HR	95% CI	P-value	HR	95% CI	P-value	HR	95% CI	P-value
Years from first visit to end-point	0.38	0.24–0.58	<0.0001	0.70	0.53–0.94	0.0174	0.60	0.38–0.95	0.0304

HR <1 favours IFNB treatment. HR, hazard ratio; CI, confidence interval; SP, secondary progression; EDSS, Expanded Disability Status Scale. Adapted from Trojano *et al.*²⁶

Cox proportional hazards regression model adjusted for PS quintiles was used to assess differences between groups of patients with early versus delayed IFNB treatment on risk of reaching a 1-point progression in the EDSS score, and the EDSS 4.0 and 6.0 milestones.⁴³ The key findings from this study are that patients who begin treatment later do not reap the same long-term benefits as those who begin treatment earlier during the disease course and that the first year from disease onset seems to represent the time frame when we could expect that initiation of an effective treatment would allow subsequent accumulation of disability to be minimized

Addressing Hidden Biases

The principal limitation of all methods of adjusting for overt biases, including PS, is the inability to address hidden biases due to unobserved or unrecorded differences between treated and control patients before treatment.³ In an ideal RCT, randomization prevents hidden biases, although even experiments may need to address some hidden biases from protocol violations, such as frequent withdrawals of patients from treatment or extensive nonadherence.⁴⁴

In a non-randomized study, the results could reflect the effects of unknown or unmeasured confounders. Hidden biases must be addressed by other means such as sensitivity analyses. A sensitivity analysis determines the magnitude of a potential unmeasured confounder that could erase the observed association or would need to be present to materially alter the conclusions of a study.^{45–46}

An example of a sensitivity analysis²⁶ performed to account for potential residual confounding caused by a hypothetical unmeasured confounder is reported in Table 5. We explored the amount of hidden bias from an unmeasured confounder necessary to alter the conclusion that RRMS patients prescribed IFNB had lower long-term disability progression than untreated patients. The results of this analysis showed that the positive effect of IFNB treatment for the time from first visit to the SP end-point (reported in the above paragraph) remained significant under high imbalanced scenarios since it might be altered by an unmeasured confounder with an HR=2 and a strong prevalence imbalance (≥80%) between the treatment group and control group (P_0-P_1) or with a low prevalence imbalance (≥20%), but an HR=8.

Table 5: Representative results of sensitivity analysis* on time from first visit to end-points SP and EDSS 4.0 and 6.0: how the magnitude of an unmeasured binary confounder might affect the propensity score-adjusted HRs of Table 4

End-point	HR [†]	P_0-P_1 [‡]	Adjusted	
			HR	95% CI
SP	2	0.8	0.66	0.41–1.00
	4	0.4	0.67	0.42–1.02
	6	0.3	0.67	0.42–1.02
	8	0.2	0.69	0.44–1.06
EDSS 4.0	2	0.1	0.76	0.58–1.03
EDSS 6.0	2	0.1	0.65	0.41–1.03

*This analysis assumes that: 1) the unmeasured confounder is binary; 2) the unmeasured confounder is independent of measured confounders; 3) there is no interaction between the unmeasured confounder and exposure; [†]Hypothetical HR of the unmeasured confounder on time to end-points; [‡]Differences in prevalence of the unmeasured confounder between IFNB-treated and controls. From Trojano *et al.*²⁶

As to end-points, EDSS scores of 4.0 and 6.0, they appeared sensitive to small bias since an unmeasured confounder with an HR=2 and a 10% prevalence imbalance would be sufficient to alter the significant effect of IFNB treatment. However, their HRs were still suggestive of a positive effect of IFNB (0.76 and 0.65, respectively). This analysis assumed that the unmeasured confounder 1) is binary, and 2) is independent of measured confounders; and that 3) no interaction occurs between the unmeasured confounder and exposure, and 4) the prevalence of the unmeasured confounder is greater in the exposed group than in the unexposed group.⁴⁶ A finding that could be sensitive to small unmeasured confounders should be interpreted with caution, but it should not be dismissed on that basis.⁴⁴ Moreover, sensitivity analysis does not demonstrate that the postulated hidden bias is necessarily present.

Conclusions

In conclusion, RCTs and observational studies do not exclude, but complement each other, and both are necessary to provide a comprehensive picture of drug benefit.⁶ All study designs have flaws that can threaten external (whether or not the study results are generalizable to other populations) or internal (how the data are gathered and assigned) validity.⁴⁷ RCTs cannot examine many different types of questions, but remain the ideal way for demonstrating a 'causal' association. However, RCTs are only as good as the quality of the randomization at baseline. If poor, correction for baseline imbalance also needs to be included, and if it is not, such studies are equally suspicious. Observational studies may be especially valuable for answering long-term questions, such as: the long-term impact of currently available disease-modifying drugs in preventing unremitting disability progression; the impact of early- versus delayed-IFNB treatment on long-term disability; and the impact of NABs on the effectiveness of IFNB in MS. Any attempt to assess treatment effectiveness within the framework of properly conducted observational studies, once overt and hidden bias are taken into account, should not have to be dismissed a priori. Methodological improvements to enhance the quality of observational studies are

essential, given the availability of large longitudinal observational data in a number of databases that are being used by MS clinicians and researchers around the world. PS-adjusted analysis, which can create groups of patients who have similar likelihood of receiving a therapy, is the most common method currently used to reduce bias in treatment comparisons in observational studies. Funding models for research should consider the substantially lower cost, the practical generalizability, and the timeliness of studies from existing databases, in comparison to RCTs.¹² It is time to give a realistic and appropriate place for observational studies in evidence-based medicine.⁴⁸

Conflicts of Interest

Maria Trojano has previously received honoraria from Sanofi-Aventis, Biogen and Bayer Schering Pharma for speaking, and research grants from Merck Serono.

Address for Correspondence

Maria Trojano, Department of Neurological and Psychiatric Sciences, University of Bari, Piazza Giulio Cesare 11 – 70124, Bari, Italy
Phone: +39 (0) 80 547 8555
Fax: +39 (0) 80 547 8532
E-mail: mtrojano@neuro.uniba.it

Received: 17 January 2009

Accepted: 25 March 2009

References

- Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity. I. Clinical trials. *Lancet* 2001; **357**: 373–380.
- D'Agostino RB Sr, Kwan H. Measuring effectiveness: what to expect without a randomized control group. *Med Care* 1995; **33**(4 suppl): AS95–AS105.
- D'Agostino RB Jr, D'Agostino RB Sr. Estimating treatment effects using observational data. *JAMA* 2007; **297**: 314–316.
- Comber H, Perry JJ. Observational studies for intervention assessment. *Lancet* 2001; **357**: 2141–2142.
- McKee M, Britton A, Black N *et al*. Methods in health services research. Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999; **319**: 312–315.
- Dobre D, van Veldhuisen DJ, DeJongste MJL *et al*. The contribution of observational studies to the knowledge of drug effectiveness in heart failure. *Br J Clin Pharmacol* 2007; **64**: 406–414.
- Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996; **312**: 1215–1218.
- MacMahon S, Collins R. Reliable assessment of the

- effects of treatment on mortality and major morbidity, II: observational studies. *Lancet* 2001; **357**: 455–462
9. Feinstein AR. *Clinical Biostatistics*, London: Mosby, 1977; pp16–20.
 10. Matthews DE, Farewell VT. *Using and Understanding Medical Statistics*, 2nd edn, revised. Basel: Karger, 1988.
 11. Rochon PA, Gurwitz JH, Sykora K *et al.* Reader's guide to critical appraisal of cohort studies: 1. Role and design. *BMJ* 2005; **330**: 895–897.
 12. Wolfe RA. Observational studies are just as effective as randomized clinical trials. *Blood Purif* 2000; **18**: 323–326.
 13. Reeves GK, Cox DR, Darby SC *et al.* Some aspects of measurement error in explanatory variables for continuous and binary regression models. *Stat Med* 1988; **17**: 2157–2177.
 14. Bar-Oz B, Moretti ME, Mareels G *et al.* Reporting bias in retrospective ascertainment of drug-induced embryopathy. *Lancet* 1999; **354**: 1700–1701.
 15. Swan SH, Shaw GM, Schulman J. Reporting and selection bias in casecontrol studies of congenital malformations. *Epidemiology* 1992; **3**: 356–363.
 16. Klein-Gelink JE, Rochon PA, Dyer S *et al.* Readers should systematically assess methods used to identify, measure and analyze confounding in observational cohort studies. *J Clin Epidemiol* 2007; **60**: 766–772.
 17. Hennekens CH, Buring JE, Manson JE *et al.* Lack of effect of longterm supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *N Engl J Med* 1996; **334**: 1145–1149.
 18. Ad Hoc Subcommittee of the Liaison Committee of the World Health Organization and the International Society of Hypertension. Effects of calcium antagonists on the risks of coronary heart disease, cancer and bleeding. *J Hypertens* 1997; **15**: 105–115.
 19. Nelson HD, Humphrey LL, Nygren P *et al.* Postmenopausal hormone replacement therapy: scientific review. *JAMA* 2002; **288**: 872–881.
 20. Normand SLT, Sykora K, Li P *et al.* Reader's guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. *BMJ* 2005; **330**: 1021–1023.
 21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
 22. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**: 516–524.
 23. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
 24. Rosenbaum PR. Propensity score. In: *Encyclopedia of Biostatistics* (Armitage P, Colton T, eds), Vol. 5 (Pri–Sph). New York: Wiley; 1998:3551–5.28.
 25. Trojano M, Russo P, Fuiani A *et al.* The Italian Multiple Sclerosis Database Network (MSDN): the risk of worsening according to IFNbeta exposure in multiple sclerosis. *Mult Scler* 2006; **12**: 578–585.
 26. Trojano M, Pellegrini F, Fuiani A *et al.* New natural history of interferon-beta-treated relapsing multiple sclerosis. *Ann Neurol* 2007; **61**: 300–306.
 27. D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a nonrandomized control group. *Stat Med* 1998; **17**: 2265–2281.
 28. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; **127**(Part 2): 757–763.
 29. Weitzen S, Lapane KL, Toledano AY *et al.* Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2004; **13**: 841–853.
 30. Shah BR, Laupacis A, Hux JE *et al.* Propensity score methods give similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58**: 550–559.
 31. Yanovitzky I, Zanutto E, Hornik R. Estimating causal effects of public health education campaigns using propensity score methodology. *Eval Program Plann* 2005; **28**: 209–220.
 32. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007; **26**: 3078–3094.
 33. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol* 2008; **61**: 537–545.
 34. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006; **98**: 253–259.
 35. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg* 2007; **134**: 1128–1135.
 36. Stenestrand U, Wallentin L, for the Swedish Register of Cardiac Intensive Care (RIKS-HIA). Early statin treatment following acute myocardial infarction and 1-year survival. *JAMA* 2001; **285**: 430–436.
 37. Gum PA, Thamilarasan M, Watanabe J *et al.* Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. *JAMA* 2001; **286**: 1187–1194.
 38. Kern LM, Powe NR, Levine MA *et al.* Association between screening for osteoporosis and the incidence of hip fracture. *Ann Intern Med* 2005; **142**: 173–181.
 39. Ko DT, Chiu M, Austin PC *et al.* Safety and effectiveness of drug-eluting stents among diabetic patients: a propensity analysis. *Am Heart J* 2008; **156**: 125–134.
 40. Paolicelli D, Lavolpe V, Pellegrini F *et al.* A post-marketing evaluation of the impact of neutralizing antibodies on the effectiveness of interferon beta in relapsing-remitting multiple sclerosis. *Mult Scler* 2008; **14**: S24; 68.
 41. Vukusic S, Confavreux C. Prognostic factors for progression of disability in the secondary progressive phase of multiple sclerosis. *J Neural Sci* 2003; **206**: 135–137.
 42. Pittock SJ, Mayr WT, McClelland RL *et al.* Change in MS related disability in a population-based cohort. A 10-year follow-up study. *Neurology* 2004; **62**: 51–59.
 43. Trojano M, Pellegrini F, Paolicelli D *et al.* Real-life impact of early interferon-beta therapy in relapsing multiple sclerosis. *Ann Neurol* 2009 (in press)
 44. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002; **137**: 693–695.
 45. Rosenbaum PR. Discussing hidden bias in observational studies. *Ann Intern Med* 1991; **115**: 901–905.
 46. Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* 1998; **54**: 948–963.
 47. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized studies. *Stat Med* 2007; **26**: 20–36.
 48. Trojano M. Is it time to recognize the use of observational data to estimate treatment effectiveness in multiple sclerosis? *Neurology* 2007; **69**: 1478–1479.