Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2008;37:1142–1147
doi:10.1093/ije/dyn079

THEORY AND METHODS

Systematic differences in treatment effect estimates between propensity score methods and logistic regression

Edwin P Martens,[1,2] Wiebe R Pestman,[2] Anthonius de Boer,[1] Svetlana V Belitser[1] and Olaf H Klungel[1]*

Accepted 1 April 2008

Background In medical research both propensity score methods and logistic regression analysis are used to estimate treatment effects in observational studies. From literature reviews it has been concluded that treatment effect estimates from both methods are quite similar. With this study we will show that there are systematic differences which can be substantial.

Methods We used a simulated population with a known marginal treatment effect and applied a propensity score method and logistic regression analysis to adjust for confounding.

Results The adjusted treatment effect in logistic regression is in general further away from the true marginal treatment effect than the adjusted effect in propensity score methods. The difference is systematic and dependent on the incidence proportion, the number of prognostic factors and the magnitude of the treatment effect. For instance, a substantial difference of 20% is found when the treatment effect is 2.0, the incidence proportion is 0.20 and there are more than 11 prognostic factors.

Conclusions Propensity score methods give in general treatment effect estimates that are closer to the true marginal treatment effect than a logistic regression model in which all confounders are modelled.

Keywords Propensity scores, confounding, adjusted treatment effect, logistic regression, conditional treatment effect, marginal treatment effect, observational studies

Introduction

A commonly used statistical method in observational studies that adjusts the estimated treatment effect for confounding, is the method of propensity scores (PS).[1,2] This method focusses on the balance of covariates between treatment groups before relating treatment to outcome. In contrast, classical methods like linear regression, logistic regression (LReg) or Cox proportional hazards regression (Cox PH) directly relate outcome to treatment and covariates by a multivariable model. In two recent literature studies it is concluded that treatment effects estimated by PS methods and regression techniques are in general fairly similar to each other.[3,4] Instead of a focus on the similarity in treatment effects between both methods, we will illustrate that the differences between PS methods and LReg analysis are systematic

[1] Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands.
[2] Centre for Biostatistics, Utrecht University, Utrecht, The Netherlands.
* Corresponding author. Department of Pharmacoepidemiology and Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Sorbonnelaan 16, 3584 CA Utrecht, The Netherlands. E-mail: o.h.klungel@uu.nl

and can be substantial. We will also demonstrate that treatment effect estimates from PS methods are in general closer to the true marginal treatment effect than the estimates from LReg analysis.

Systematic differences between treatment effect estimates

In the literature review of Shah *et al*., the main conclusion was that PS methods resulted in similar treatment effects compared to traditional regression modelling.[3] This was based on the agreement that existed between the significance of treatment effect in PS methods compared to LReg or Cox PH methods in 78 reported analyses. This agreement was denoted as excellent ($\kappa = 0.79$) and the mean difference in the logarithm of treatment effects was quantified as 6.4%. In the review of Stürmer *et al*. it was also stressed that PS methods did not result in substantially different treatment effect estimates compared to LReg or Cox PH methods.[4] They reported that in only 9 out of 69 studies (13%) the effect estimate differed by more than 20%.

The results of these reviews can also be interpreted differently: the dissimilarity between methods is systematic resulting in treatment effect estimates that are on average stronger in LReg and Cox PH analysis. In Shah *et al*. the disagreement between methods was in the same direction: all eight studies that disagreed resulted in a significant effect in LReg or Cox PH methods and a nonsignificant effect in PS methods ($P = 0.008$, McNemar's test). Similarly, the treatment effect in PS methods was more often closer to unity than in LReg or Cox PH (34 versus 15 times, $P = 0.009$, binomial test with $\pi_0 = 0.5$). In the review of Stürmer *et al*., it turned out that substantial differences between both methods only existed when the estimates in LReg or Cox PH were *larger* than in PS methods.

For a more complete view we combined the results of these two literature reviews by including all studies that reported treatment effects for PS methods (matching, stratification or covariate adjustment) and regression methods (LReg or Cox PH) and by excluding studies that were found in both reviews. From all 96 studies there were twice as many studies in which the treatment effect from LReg or Cox PH methods was larger than from PS methods: 50 versus 24 (=68%). Testing the null hypothesis of equal proportions (binomial test, $\pi_0 = 0.5$) resulted in a highly significant difference ($P = 0.003$). The mean difference in the logarithm of treatment effects between both methods (δ)[3] was calculated at 5.0%, significantly different from 0 ($P = 0.001$, 95% confidence interval (CI): 2.0, 7.9). In the 21 studies with treatment effects larger than an odds ratio (OR) of 2.0 or smaller than 0.5 this mean difference was considerably larger: $\delta = 19.0\%$, 95% CI: 10.3, 27.6.

From these literature reviews we conclude that PS methods significantly more often result in treatment effects closer to the null hypothesis of no effect than LReg or Cox PH methods. The larger the treatment effects, the larger the differences.

Explaining the differences in treatment effect estimates

The reason for the systematic differences between treatment effect estimates from PS methods and LReg or Cox PH methods can be found in the concept of *non-collapsibility* of the OR and hazard ratio. In the literature this phenomenon has been recognized and described by many authors.[5–15] Although this concept has different interpretations and is sometimes used as a synonym for *confounding*, we will distinguish between both concepts.[11] The concept of non-collapsibility can be easiest understood in the absence of confounding, where confounding is defined as bias in estimating the true treatment effect due to differences in the distributions of prognostic factors between treatment groups.

When two populations exist that are similar on all prognostic factors (for instance one treated and one untreated), these populations are called *exchangeable*. In such situations a *marginal treatment effect* can be estimated, on average similar to the treatment effect in randomized studies: this is the effect of treating a certain population instead of not treating that population. Note that this treatment effect is defined by only relating treatment to outcome, without using covariates. That means that it can be estimated without adjustment for covariates by an *unadjusted treatment effect*, for instance by a difference in means, a risk ratio or an OR. When on the other hand both populations are not similar on prognostic factors, as is to be expected in observational studies, one should estimate an *adjusted treatment effect*, trying to adjust for all potential confounders. This can be done for instance by any multivariable regression model or by one of the PS methods. The type of treatment effect as will be estimated when covariates are involved, is also called a *conditional treatment effect* or *subject-specific effect*.[9]

When treated and untreated populations are exactly similar on all covariates, unadjusted and adjusted treatment effects should coincide, because the primary objective of adjustment is to adjust for dissimilarities in covariate distributions: if there are none, ideally adjustment should have no effect. Unfortunately, this is not generally true, for instance when ORs from LReg analysis are used to quantify treatment effects. Consider two LReg models:

$$\text{logit}(y) = \alpha_1 + \beta_t t \quad (1)$$

$$\text{logit}(y) = \alpha_2 + \beta_t^* t + \beta_1 x_1 \quad (2)$$

where y is a dichotomous outcome, t a dichotomous treatment, x_1 a dichotomous prognostic factor and α_1 and α_2 constants, e^{β_t} the unadjusted treatment effect, $e^{\beta_t^*}$ the adjusted treatment effect and e^{β_1} the effect of x_1.

Suppose that in a certain situation only one prognostic factor exists (x_1) with a different distribution for both treatment groups. An adjusted treatment effect β_t^* will in general be interpreted as an estimate for the true marginal treatment effect, i.e. the effect that would be found when both treatment groups had *similar* distributions of x_1. But when in reality the distribution of x_1 is similar for both treatment groups and model 2 is applied, it turns out that the adjusted treatment effect estimate β_t^* does not equal the unadjusted treatment effect β_t. More generally, when both treatment groups are similar with respect to their covariate distributions, the adjusted and unadjusted treatment effects will not coincide in non-linear regression models or generalized linear models with another link function than the identity link (equalling a linear regression analysis) or log-link. We refer to the literature for a mathematical explanation of this phenomenon[5,6,16] and will illustrate in the next paragraph its implications for the comparison between LReg and PS methods in epidemiological research.

Adjusting for equally distributed prognostic factors

To illustrate the non-collapsibility of the OR, we created a data set of $n = 100\,000$, a binary out-come y (π_y varying from 0.02 to 0.20), a treatment t ($\pi_t = 0.50$) and 20 binary prognostic factors $x_1, \ldots, x_{20}$ with $\pi_{x_1} = \ldots = \pi_{x_{20}} = 0.50$ and $e^{\beta_{x_1}} = \ldots = e^{\beta_{x_{20}}} = 2.0$. These factors, which we will call *non-confounders*, were exactly equally distributed across treatments $t = 1$ and $t = 0$. The true average treatment effect is therefore known and equals the unadjusted effect of treatment on outcome e^{β_t} in Equation 1, which was set to 2.0. First we included the factor x_1 in the LReg model of Equation 2 and calculated an adjusted treatment effect $e^{\beta_t^*}$. We extended this model by including the factors from x_2 to x_{20} and calculated the corresponding adjusted treatment effects. Because this data set is only created to illustrate a mathematical result, sampling is not necessary here.

In Figure 1 all these adjusted treatment effects were plotted for various incidence proportions. For example, with an incidence proportion of $\pi_y = 0.10$ the adjusted treatment effect is estimated as nearly 2.16 in a LReg model with 10 non-confounders and as 2.43 in a model with 20 non-confounders. Its increase is stronger when the incidence proportion is higher. Also an increase in the strength of the treatment effect (here fixed at 2.0) or an increase in the strength of the association between non-confounders and outcome (also fixed at 2.0) will increase the difference between adjusted and unadjusted treatment effect estimates (data not shown).[17]

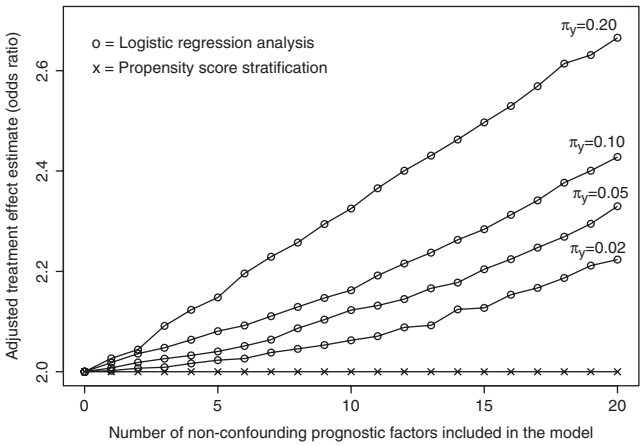

Figure 1 Adjusted treatment effects for 1–20 non-confounding prognostic factors and various incidence proportions in LReg and PS stratification ($n = 100\,000$, $e^{\beta_t} = 2.0$)

This is in sharp contrast with PS methods for which treatment effects remain unchanged, irrespective of the number of covariates in the PS model, the incidence proportion, the strength of the treatment effect and the strength of the association between non-confounders and outcome. The reason is that all prognostic factors are equally distributed between treatment groups (univariate as well as multivariate), which means that the calculated PS is constant for every individual. Stratification on the PS or including the PS as a covariate will leave the unadjusted treatment effect unchanged. Although it seems obvious, it illustrates an important difference between PS methods and LReg: PS methods concentrate on similarity of groups, while LReg analyses model the outcome. Because the treatment effect for both analyses is the OR, it will cause differences between the effect estimates.

Adjusting for imbalanced prognostic factors

Perfectly balanced treatment groups, as used in the previous paragraph, are quite exceptional in practice. In general, treatment groups will differ from each other with respect to covariate distributions, in observational studies (systematic and random imbalances), but also in randomized studies (random imbalances). In this paragraph, we will explore the differences between LReg and PS analysis when adjustment takes place for imbalanced prognostic factors. In simulation studies it is common to create imbalance between treatment groups by first modelling treatment as a function of covariates and then outcome as a function of treatment and covariates.[18–21] Unfortunately, the treatment effect that is defined in such studies as the

true treatment effect is an *adjusted* treatment effect which is conditional on the covariates that has been chosen in the true model. One solution is to calculate such a true treatment effect with an iterative procedure,[22] but still all data are based on LReg models, one of the methods to be evaluated.[23] These problems can be circumvented when one starts with a balanced population with a known treatment effect in which no outcome model is involved in generating the data. By using the imbalances on prognostic factors that appear in random samples, the effects of adjustment between LReg and PS methods can be compared. Random imbalances are indistinguishable from systematic model-based imbalances at the level of an individual data set: they only differ from one another by the fact that random imbalances will cancel out when averaged over many samples. For illustrating the differences between LReg and PS methods when adjusting for imbalances it is not important *how* imbalances have arisen.

Simulations

We created a population of $n = 100\,000$, a binary outcome y ($\pi_y = 0.30$), treatment t ($\pi_t = 0.50$) and five normally distributed prognostic factors $x_1, \ldots, x_5$ with mean = 0.50, standard deviation = 0.4 and $e^{\beta_{x_1}} = \ldots = e^{\beta_{x_5}} = 2.0$. The true marginal treatment effect in the population was set to $e^{\beta_t} = 2.5$. To randomly create imbalances, we took 1000 random samples with varying sample sizes ($n = 200$, 400, 800 and 1600). The LReg model used for adjustment is:

$$\text{logit}(y) = \alpha_y + \beta_t^* t + \beta_{1y} x_1 + \cdots + \beta_{5y} x_5 \qquad (3)$$

and the PS are calculated as:

$$\text{PS} = \frac{e^{\text{logit}(t)}}{1 + e^{\text{logit}(t)}} \qquad (4)$$

with $\text{logit}(t) = \alpha_t + \beta_{1\,t} x_1 + \ldots + \beta_{5t} x_5$.

To adjust for confounding we stratified subjects on the quintiles of the PS and calculated a common treatment effect using the Mantel-Haenszel estimator.

Comparison of adjusted treatment effects

For all 1000 samples adjusted treatment effects were estimated by LReg analysis (Equation 3) and PS stratification with five strata based on the quintiles of the PS (Equation 4). In Figure 2, density estimation[24,25] is used to illustrate that the ratio of adjusted treatment effects in a LReg analysis with $n = 400$ are nearly 9% larger than those in PS stratification: in almost all samples (97%) this ratio is larger than one, indicating larger effects for LReg analysis. This confirms the results found in the two literature reviews that LReg or Cox PH result in general in higher treatment effects than PS analysis (50/74 = 68%). The difference between both percentages is due to the diversity in models, treatment effects, sample sizes and number of confounders that were found in the literature.

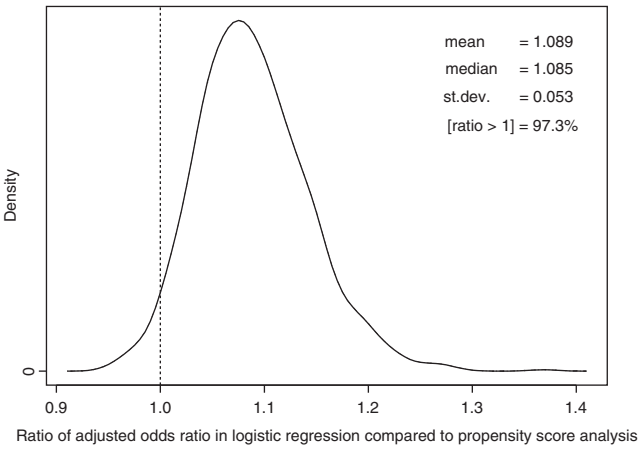

Figure 2 Density function of the ratio of adjusted odds ratios of treatment effect in LReg compared with PS analysis, 1000 samples of $n = 400$

Table 1 Summary measures of the ratio of adjusted odds ratios of treatment effect in LReg compared with PS analysis in 1000 samples

	$n = 200$	$n = 400$	$n = 800$	$n = 1600$
Mean	1.102	1.087	1.085	1.082
Median	1.094	1.081	1.082	1.082
Standard deviation	0.096	0.055	0.038	0.030
Fraction > 1	0.887	0.970	0.994	0.999

In Table 1, the results are summarized for various sample sizes. Between sample sizes of 400, 800 and 1600 there are only minor differences in the mean and median ratio. Overall it can be concluded that with the chosen associations and number of covariates, the adjusted treatment effect in LReg is 8–10% higher than in PS analysis, slightly decreasing with sample size. We also varied the number of strata (5, 7 or 10), which has in this setting only negligible effects on the results.

Comparison of adjusted and unadjusted treatment effects

Apart from a comparison between LReg and PS methods, it is relevant to compare the adjusted effect in both methods to the unadjusted, marginal effect. Ideally, the average of the ratio of adjusted to unadjusted effect should be located around one, because then the adjusted effect is an unbiased estimator for the true treatment effect.

The results are presented in Figure 3 for sample sizes of 400. It can be seen that the distribution for PS stratification is centered around 1.0, while the distribution for LReg analysis is located further to the right. This means that when the adjusted, conditional

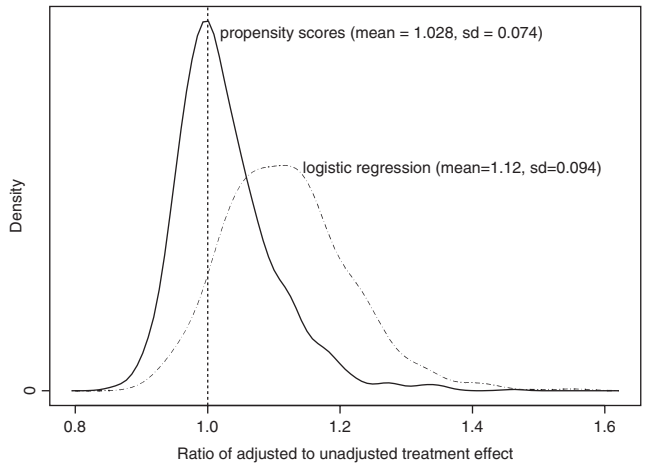

Figure 3 Density functions of the ratio of adjusted to unadjusted odds ratios of treatment effect in LReg and PS analysis, 1000 samples of $n = 400$

treatment effect is used instead of the unadjusted treatment effect (which equals in this setting on average the true marginal effect), LReg analysis overestimates this effect on average with 12%. In contrast, PS stratification only results on average in an overestimation of less than 3%. Another difference is the smaller standard deviation in PS analysis (0.074) compared to LReg (0.094). When the number of prognostic factors, the incidence proportion, the strength of the treatment effect or the strength of the association between prognostic factors and outcome is increased, the overestimation in LReg compared to PS methods will also increase.[17]

Conclusion and discussion

In medical studies, LReg analysis and PS methods are both applied to estimate an adjusted treatment effect in observational studies. In the literature, the effect estimates of both methods are often classified as 'similar' and 'not substantially different'. We pointed out that the differences between the methods are due to the use of the OR as treatment effect and the non-collapsibility of this measure. The differences between PS methods and LReg analysis are systematic and can be substantial, especially when the number of prognostic factors is more than five, the treatment effect is larger than an OR of 1.25 (or smaller than 0.8) or the incidence proportion is between 0.05 and 0.95. This difference is frequently overlooked by analysts in the literature. With respect to the objective to adjust for the imbalance of covariate distributions between treatment groups, we illustrated that the estimate of PS methods is in general closer to the true marginal treatment effect than the estimate of LReg analysis.

We showed that the number of included factors in the outcome model is one of the explanations for the difference in treatment effect estimates between the studied methods in which ORs are involved. For PS methods without further adjustment, this is only two (i.e. the PS and treatment), while for LReg this is in general much larger (the number of included covariates plus one). For that reason it is to be expected that the main results are not largely dependent on the specific PS method used (stratification, matching, covariate adjustment or weighting), except when PS methods are combined with further adjustment for confounding by entering some or all covariates separately in the outcome model. Besides PS stratification we also used covariate adjustment using the PS. We hardly found any differences and speculate that the same is true for other PS methods like matching or weighing on the PS.

We used only the most simple PS model (all covariates linearly included) and did not make any effort to improve the PS model in order to minimize imbalances.[26] When a more optimal PS model has been chosen, it is to be expected that the PS estimate will be closer to the true marginal treatment effect than reported above.

When two alternative adjustment methods give different results, naturally the question arises which one to use. In the literature one will often find the argument that physicians are more interested in conditional effects because they have to make appropriate treatment decisions for individual patients, while marginal effects are more suitable for epidemiologists or policy makers who want to know the treatment effect for a specified population.[18] It is true that in clinical practice decisions are made for individual patients, but the conditional effect estimated in a LReg analysis is still averaged over subgroups. Furthermore, individual treatment decisions can be better made on absolute risk reductions instead of relative risks or ORs.[27] Also, when we review the literature on treatment effects it is clear that most authors have a marginal effect in mind, the effect that would have been found when it was possible to do a randomized study.[3,4,28–31] These arguments will favour PS methods as adjustment method above LReg analysis, because the former is in general closer to the marginal treatment effect.

References

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

[2] D'Agostino RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;**17:**2265–81.

[3] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005;**58:**550–59.

[4] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol* 2006;**59:**437–47.

[5] Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984;**71:**431–44. Available at: http://www.jstor.org/stable/view/2336553.

[6] Gail MH. The effect of pooling across strata in perfectly balanced studies. *Biometrics* 1988;**44:**151–63.

[7] Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991;**58:**227–40.

[8] Guo J, Geng Z. Collapsibility of logistic regression coefficients. *J R Statist Soc B* 1995;**57:**263–67.

[9] Hauck WW, Neuhaus JM, Kalbfleisch JD *et al*. A consequence of omitted covariates when estimating odds ratios. *J Clin Epidemiol* 1991;**44:**77–81.

[10] Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials* 1998;**19:**249–56.

[11] Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Stat Science* 1999;**14:**29–46.

[12] Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987;**125:**761–68.

[13] Wickramaratne PJ, Holford ThR. Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. *Biometrics* 1987;**43:**751–65, [Erratum, Biometrics 1989; 45:1039].

[14] Bretagnolle J, Huber-Carol C. Effects of omitting covariates in Coxs model for survival data. *Scand J Stat* 1988;**15:**125–38.

[15] Morgan TM, Lagakos SW, Schoenfeld DA. Omitting covariates from the proportional hazards model. *Biometrics* 1986;**42:**993–95.

[16] Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Statist-Theory Meth* 1991;**20:**2609–31.

[17] Rosenbaum PR. Propensity score. In: Armitage P, Colton T (eds). *Encyclopedia of Biostatistics*. Chichester, United Kingdom: Wiley, 1998.

[18] Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med* 2007;**26:**754–68.

[19] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007; **26:**734–53.

[20] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;**158:**280–87.

[21] Negassa A, Hanley JA. The effect of omitted covariates on confidence interval and study power in binary outcome analysis: A simulation study. *Cont Clin trials* 2007;**28:** 242–48.

[22] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007;**26:**3078–94.

[23] Martens EP, Pestman WR, Klungel OH. Letter to the editor: 'Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study, by Austin PC, Grootendorst P, Normand ST, Anderson GM'. *Stat Med* 2007;**26:**3208–10.

[24] Silverman BW. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[25] Wand MP, Jones MC. *Kernel Smooting*. London, UK: Chapman and Hall, 1995.

[26] Rubin DB. On principles for modeling propensity scores in medical research (Editorial). *Pharmacoepidemiol Drug Saf* 2004;**13:**855–57.

[27] Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;**365:**176–86.

[28] Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000;**342:**1878–86.

[29] Ioannidis JP, Haidich AB, Pappa M *et al*. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA* 2001;**286:**821–30.

[30] Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and nonrandomised clinical trials. *BMJ* 1998;**317:**1185–90.

[31] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342:**1887–92.