3. Hotelling H. Book review: the triumph of mediocrity in business. *Journal of American Statistical Association* 1933; **28**:463–465.

4. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *Journal of Chronic Diseases* 1962; **15**:969–977.

5. Galton F. Regression toward mediocrity in hereditary stature. *Journal of Anthropological Institute of Great Britain and Ireland* 1886; **15**:246–263.

Published online 25 October 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2722

LETTER TO THE EDITOR

Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study (p n/a)

by Peter C. Austin, Paul Grootendorst, Sharon-Lise T. Normand, Geoffrey M. Anderson, *Statistics in Medicine*, Published Online: 16 June 2006, DOI: 10.1002/sim.2618

From: *Edwin P. Martens*[1,2], *Wiebe R Pestman*[2] *and Olaf H. Klungel*[1]
[1]*Department of Pharmacoepidemiology and Pharmacotherapy*, *Utrecht University*, *Utrecht*, *The Netherlands*
[2]*Centre for Biostatistics*, *Utrecht University*, *Utrecht*, *The Netherlands*

In a recent simulation study Austin *et al*. conclude that conditioning on the propensity score gives biased estimates of the true conditional odds ratio of treatment effect in logistic regression analysis. Although we generally agree with this conclusion, it can be easily misinterpreted because of the word bias. From the same study one can similarly conclude that logistic regression analysis will give a biased estimate of the treatment effect that is estimated in a propensity score analysis. Because propensity score methods aim at estimating a marginal treatment effect, we believe that the last statement is more meaningful.

DIFFERENT TREATMENT EFFECTS

The authors raise an important issue, which is probably unknown to many researchers, that in logistic regression analysis a summary measure of conditional treatment effects will in general not be equal to the marginal treatment effect. This phenomenon is also known as non-collapsibility of the odds ratio [1], but is apparent in all non-linear regression models and generalized linear models with a link function other than the identity link (linear models) or log-link function [2]. In other words, even if a prognostic factor is equally spread over treatment groups, the inclusion of this variable in a logistic regression model will increase the estimated treatment effect. This increasing effect of a conditional treatment effect compared to the overall marginal effect is larger when more

Copyright © 2006 John Wiley & Sons, Ltd.

prognostic factors are added, but lower when the treatment effect is closer to $\mathrm{OR}=1$ and also lower when the incidence rate of the outcome is smaller [3]. In general, it can be concluded that in a given research situation many different conditional treatment effects exist, depending on the number of prognostic factors in the model.

TRUE CONDITIONAL TREATMENT EFFECT

The true treatment effect is the effect on a specific outcome of treating a certain population compared to not treating this population. In randomized studies this can be estimated as the effect of the treated group compared to the non-treated group. The true conditional treatment effect as defined in Austin *et al.* is the treatment effect in a certain population given the set of six prognostic factors and given that the relationships in the population can be captured by a logistic regression model. Two of the six prognostic factors were equally distributed between treatment groups and *included* in the equation for generating the data. But there are also non-confounding prognostic factors *excluded* from this equation, because not all of the variation in the outcome is captured by the six prognostic factors. That means that it seems to be at least arbitrary how many and which of the non-confounding prognostic factors were included or excluded to come to a 'true conditional treatment effect'. Because of the non-collapsibility of the odds ratio, all these conditional treatment effects are in general different from each other, but which of these is the one of interest remains unclear. The only thing that is clear, is that application of the model that was used to generate the data will find on average this 'true conditional treatment effect', while all other models, including less or more prognostic factors, will in general find a 'biased' treatment effect. It should be therefore no surprise that propensity score models will produce on average attenuated treatment effects, for propensity score models correct for only one prognostic factor, the propensity score. This implies that the treatment effect estimates from propensity score models are in principal closer to the overall marginal treatment effect than to one of the many possible conditional treatment effects.

MARGINAL OR CONDITIONAL TREATMENT EFFECTS?

The authors give two motivations why a conditional treatment effect is more interesting than the overall marginal treatment effect (which is the effect that would be found if treatments were randomized). Firstly, they indicate that a conditional treatment effect is more interesting to physicians, because it allows physicians to make appropriate treatment effect decisions for specific patients. Indeed, in clinical practice treatment decisions are made for individual patients, but these decisions are better informed by subgroup analyses with specific treatment effects for subgroups: a specific conditional treatment effect is still some kind of 'average' over all treatment effects in subgroups. Another argument is that treatment decisions on individual patients should be based on the absolute risk reduction and not on odds ratios or relative risks [4]. Secondly, the authors suggest that in practice researchers use propensity scores for estimating conditional treatment effects. However, in most studies in which propensity scores and logistic regression analysis are both performed, researchers rather have an overall marginal treatment effect in mind than one specific conditional treatment effect [5]. Furthermore, the overall marginal treatment effect is one well-defined treatment effect, whereas conditional treatment effects are effects that are dependent on the chosen model. The reason for comparing propensity score methods with logistic regression

Copyright © 2006 John Wiley & Sons, Ltd.

analysis is probably not because the aim is to estimate conditional effects, but simply because logistic regression is the standard way of estimating an adjusted treatment effect when the outcome is dichotomous.

In conclusion, propensity score methods aim to estimate a marginal effect, which in general is not a good estimate of a conditional effect in logistic regression analysis because of the non-collapsibility of the odds ratio. An overall marginal treatment effect is better defined and seems to be of more interest than all possible conditional treatment effects. Finally, these conditional effects are dependent on the number of non-confounders, which is not the case for propensity score methods.

REFERENCES

1. Greenland S, Robins MR, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**:29–46.
2. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **71**:431–444.
3. Rosenbaum PR. Propensity score. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). Wiley: Chichester, U.K., 1998.
4. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. From subgroups to individuals: general principles and the example of carotid endarterectomy. *The Lancet* 2005; **365**:256–265.
5. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 2005; **58**(6):550–559.

Published online 20 March 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2878

AUTHORS' REPLY

We would like to thank Professors Martens, Pestman, and Klungel (henceforth MPK) for their letter to the editor concerning our recent article in *Statistics in Medicine* [1]. The article addressed by MPK is the most recent in a series of articles that we have published examining different properties of propensity methods [1–4]. In the article addressed by MPK, we described two different measures of treatment effect: marginal or average treatment effects and conditional or adjusted treatment effects. Researchers using observational or non-randomized data may be less familiar with the existence of marginal or average treatment effects. Historically, regression adjustment has been used to estimate the effects of exposures or interventions when using observational data. However, there is an increasing interest in using propensity score methods to estimate the effects of exposures or interventions when using observational data. Frequently, authors report treatment effects obtained using both regression adjustment and propensity score methods [5], where one analysis is presumably used as a confirmatory analysis. The objective of our recently published article was to illustrate that, except in certain restrictive circumstances, the use of propensity score methods results in biased estimation of conditional or adjusted treatment effects.

The primary objection of MPK is that one could have similarly concluded that logistic regression analysis will give a biased estimate of the treatment effect that is estimated in a propensity score analysis. While this result could be shown, it does not follow directly from our published study.

Copyright © 2006 John Wiley & Sons, Ltd.