European Heart Journal Advance Access published June 28, 2012

European Heart Journal
doi:10.1093/eurheartj/ehs186

EDITORIAL

Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting

Gary S. Collins[1]* and Yannick Le Manach[2]

[1]Centre for Statistics in Medicine, Wolfson College Annexe, University of Oxford, Linton Road, Oxford OX2 6UD, UK; and [2]Department of Anesthesiology, and Critical Care, Centre Hospitalo-Universitaire Pitié-Salpêtrière, Paris Cedex 13, France

This editorial refers to 'Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndrome', by I.J. Dahabreh *et al.*, doi:10.1093/eurheartj/ehs114

Evaluating the effectiveness of a therapeutic intervention is ideally carried out in the setting of a randomized controlled trial (RCT). Patients are randomly allocated to the experimental and control groups ensuring that observed, pre-treatment key prognostic characteristics, but also unobserved patient characteristics, are balanced between the treatment groups, minimizing the variability in patient characteristics. Providing a sufficient number of patients have been randomized, this balance in observed and unobserved pre-treatment characteristics between the groups enables unbiased conclusions about the treatment effect to be drawn. There may, however, be instances where randomization is not possible due to, for example, ethical reasons (e.g. emergency surgery,[1] transplantation[2]) or because it is impractical (e.g. rare events, financial reasons). When it is not possible to conduct an adequately powered RCT, observational studies are often carried out to examine and infer treatment effects. In addition, treatment effects observed in RCTs that involve highly selective populations are often examined in different patient populations and settings in observational studies. However, in observational studies, investigators have no control over treatment assignment, which is often part of a patient's routine medical care. In these instances, it is likely that potentially large systematic differences (typically confounding by indication) in observed patient characteristics could lead to large, biased, and ultimately misleading estimates of treatment effect.

Propensity scores are increasingly being used to reduce the impact of any imbalance in pre-treatment patient characteristics and, more importantly, confounding; patient characteristics that influence treatment selection.[3] In light of the increasing number of studies, a recent article in the *European Heart Journal* provided an overview of the objectives of and approaches to propensity score analyses.[4] To summarize briefly, the propensity score of a patient is defined as the probability of receiving the experimental treatment conditional on the patient's pre-treatment characteristics. The propensity score is a multivariable model (typically using a logistic regression model) where pre-treatment characteristics (and all known potential confounders) are included in the model as predictors where the outcome is the treatment group. The propensity score (i.e. the probability of being treated given the observed pre-treatment characteristics) can then be used in a number of ways, including matching, stratification, or regression adjustment.[4] Matching and stratification are generally preferred over regression adjustment by creating a quasi-randomized study design, whereby two patients, one in each group, who have the same propensity can be assumed to have been equally likely to have been randomly allocated to each group.[3] However, unlike RCTs, the balance in unobserved pre-treatment patient characteristics (hidden bias) remains problematic for propensity scores, though the magnitude of hidden bias can be evaluated in sensitivity analyses.

Dahabreh and colleagues describe the results from an interesting study to examine the agreement in treatment effects between observational studies using propensity scores and RCTs evaluating therapeutic interventions for acute coronary syndrome.[5] By matching 21 observational studies to 63 RCTs, the authors examined the similarity of treatment effects of 17 short- and long-term outcomes. The authors highlighted that treatment effects from the observational studies were more often extreme in magnitude but rarely statistically significantly different from those reported in RCTs. We discuss and elaborate findings and methodological implications from this study.

Whilst Dahabreh and colleagues reported that treatment effects were often slightly larger in magnitude in the propensity score studies compared with those in the RCTs, they concluded there was good agreement between the treatment effects. This finding of similarity of treatment effects from propensity scores and

The opinions expressed in this article are not necessarily those of the Editors of the *European Heart Journal* or of the European Society of Cardiology.

* Corresponding author. Tel: +44 1865 284418, Fax: +44 1865 284424, Email: gary.collins@csm.ox.ac.uk

RCTs is consistent with existing studies.[6] However, publication bias and selective reporting could be an unacknowledged source of bias in the observational studies; it is likely that selective reporting of significant outcomes will contribute and be a major source of bias.[7] Whereas RCTs with clinically small treatment effects are likely to be published, it is often unlikely that similarly small treatment effects found in observational studies will be submitted for publication or indeed published. Ideally, we believe observational studies should be registered (and, where possible, published) so all outcomes, data handling, and statistical analyses are pre-specified to minimize the scope for selective outcome reporting.[8] Another interesting aspect contributing to the difference in magnitude of treatment effects, whilst not statistically different, is the issue of intention-to-treat analyses usually conducted in RCTs compared with the as-treated analyses conducted in the propensity score analyses. The authors did not dwell on this aspect, and details on how many patients in the individuals trials that received the alternative treatment to which they were randomized were not examined. Furthermore, unpicking intention-to-treat analyses and identifying analysed patients from RCTs is potentially non-trivial due to poor and inconsistent reporting.[9]

An additional important feature of the study by Dahabreh and colleagues, which could contribute to the study findings, yet only received brief attention, is the poor methodological conduct and reporting of observational studies using propensity scores. These findings in particular, which are consistent with existing systematic reviews of propensity scores, deserve a much more detailed examination.[10–12] In particular, issues of evaluating balance in pre-treatment patient characteristics and analysis strategy are two aspects that we will discuss further.

For observational studies that apply propensity scores in a matching or stratification framework, Dahabreh and colleagues, in agreement with existing systematic reviews, reported few studies assessing the balance in pre-treatment patient characteristics between treatment groups. A key component in the matching framework is to ensure that matching on the propensity score yields two groups of patients (experimental and control groups) with a similar distribution of pre-treatment patient characteristics. Balance in pre-treatment characteristics should ideally be assessed not by significance testing,[13] but by calculating standardized differences for each characteristic.[4] Any imbalance would necessitate refining the propensity score model in an iterative process of model building an assessment of balance.[14] It is also believed that calculating treatment differences in studies using propensity scores should also account for matching, yet this is either rarely done or indeed reported. Inadequate methodological rigour could contribute to inflated treatment estimates.

Whilst not mentioned by Dahabreh and colleagues, the reporting and handling of missing data in observational studies using propensity scores in general has received little attention.[15] Developing propensity models requires complete data on all predictors included in the model, which for creating a propensity model can be large, yet very few clinical data sets have complete information on all predictors for patients. Authors are encouraged to consider how missing data could affect how the propensity score was derived and ultimately how and to what extent this affects the estimation of the treatment effect.

Describe how the propensity score model was derived
Describe how missing data were handled, including
(a) report the number missing values for predictors
(b) report the number of cases with complete data
Describe how the propensity score-matched sample was created (i.e. the method of matching).
Describe how the balance between treatment groups was assessed and what was done for any predictors that were not balanced
Describe the statistical methods used to estimate the treatment effect

Figure 1 Minimal considerations for the transparent reporting of observational studies using propensity scores.

Fundamental to the validity and reproducibility of estimating treatment effects in observational studies incorporating propensity scores is clear and transparent reporting. Reporting guidelines exist for a variety of study designs and statistical analyses,[16] yet currently there are no consensus-based guidelines on either the conduct or reporting of propensity scores. As an absolute minimum we recommend that authors adhere to STROBE guidelines for reporting observational studies,[17] but, in addition, authors should be encouraged to provide sufficient and clear information on: how the propensity score was developed, the handling of missing data, how was the propensity score-matched sample created, how balance between treatment groups was assessed, and the statistical methods used to estimate the treatment effect (*Figure 1*).

Conflict of interest: none declared.

References

1. Ergina PL, Cook JA, Blazeby JM, Boutron I, Clavien PA, Reeves BC, Seiler CM; Balliol Collaboration. Challenges in evaluating surgical innovation. *Lancet* 2009;**374**:1097–104.
2. Hunt S. A fair way of donating hearts for transplantation. *BMJ* 2000;**321**:526.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
4. Heinze G, Juni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J* 2011;**32**:1704–1708.
5. Dahabreh I, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, Rassen JA, Trikalinos TA, Kitsios GD. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndrome. *Eur Heart J*. Advance Access published June 17, 2012, doi: 10.1093/eurheartj/ehs114.
6. Kuss O, Legler T, Borgermann J. Treatments effects from randomized trials and propensity score analyses were similar in similar populations in an example from cardiac surgery. *J Clin Epidemiol* 2011;**64**:1076–1084.
7. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;**337**:867–872.
8. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? *CMAJ* 2010;**182**:1638–1642.
9. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ* 1999;**319**:670.
10. 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.

11. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2004;**13**:841–853.
12. Sturner 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**: 437–447.
13. Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;**335**:149–153.
14. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;**79**:516–524.
15. D'Agostino RB Jr, Rubin DB. Estimating and using propensity scores with partially missing data. *J Am Stat Assoc* 2000;**95**:749–59.
16. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. *BMC Med* 2010;**8**:24.
17. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;**335**:806.