SPECIAL REPORT

Journal of Comparative Effectiveness Research

Propensity score models in observational comparative effectiveness studies: cornerstone of design or statistical afterthought?

John W Robinson

John W Robinson, MD, PhD, LLC,
Statistical & Health Informatics Consulting,
4303 Stanford Street, Chevy Chase, MD 20815, USA
Tel.: +1 301 652 3579
Fax: +1 301 652 0599
john@medicalcareanalytics.com

Propensity score models are increasingly used in observational comparative effectiveness studies to reduce confounding by covariates that are associated with both a study outcome and treatment choice. Any such potentially confounding covariate will bias estimation of the effect of treatment on the outcome, unless the distribution of that covariate is well-balanced between treatment and control groups. Constructing a subsample of treated and control subjects who are matched on estimated propensity scores is a means of achieving such balance for covariates that are included in the propensity score model. If, during study design, investigators assemble a comprehensive inventory of known and suspected potentially confounding covariates, examination of how well this inventory is covered by the chosen dataset yields an assessment of the extent of bias reduction that is possible by matching on estimated propensity scores. These considerations are explored by examining the designs of three recently published comparative effectiveness studies.

Keywords: comparative effectiveness research ▪ confounding ▪ observational study ▪ propensity score ▪ selection bias

Propensity score models are increasingly used in observational comparative effectiveness studies to reduce confounding by covariates that are associated with both a study outcome and treatment choice, a major threat to internal validity. Given the important role that comparative effectiveness research is expected to play in improving the quality and efficiency of healthcare services, propensity score models are likely to become an increasingly prevalent feature of the methodological landscape. Thus, it is useful to consider how these models can be used to their fullest potential; not simply as an analytical tool, but as a guiding structure during study design [1,2].

The propensity score is most easily considered from the perspective of a randomized trial or observational cohort study with a binary treatment variable, in which patients are assigned to (or choose) either the treatment or control condition. From this perspective, the propensity score is the probability of assignment to (or choice of) the treatment condition, given the values of observed covariates – characteristics of the patient, provider and treatment context measured prior to treatment assignment or choice [1–4]. In a randomized trial, propensity scores are fixed by design, for example, at 0.5 for each subject in a trial with equal allocation between treatment and control groups. In an observational study, propensity scores cannot be known, but they can be estimated [1–4].

Propensity scores for a binary treatment choice can be estimated by fitting a logistic regression model to the full study sample, with the binary treatment variable as response and observed covariates as predictors [2–4]. The fitted model yields an estimated propensity score for each subject. Then, if a subsample is formed by matching

Future Medicine part of fsg

each treated subject with a fixed number of controls with very similar estimated propensity scores (e.g., one control for 1:1 matching), the distributions of observed covariates will tend to be well balanced between treatment and control groups in this matched subsample, akin to the result achieved by random assignment [2–4]. Of course, it should be noted that whereas the balancing effect of randomization applies equally to observed and unobserved (unmeasured) covariates, matching on estimated propensity scores cannot be expected to balance covariates that are unobserved and thus unavailable for inclusion in the propensity score model. Also, it should be noted that some treated subjects may be unmatchable due to absence of an available control with a similar estimated propensity score (due, in turn, to absence of an available control with similar values of potentially confounding covariates), and some controls may be unmatchable for analogous reasons; hence the matched subsample is generally a subset of the full study sample.

A beneficial consequence of this matching approach is that covariate balance in the propensity-matched subsample can be evaluated using a tabular format, analogous to the typical 'Table 1' from a randomized trial that compares covariate distributions between treatment and control groups. Quantitative assessment of balance can be accomplished by comparing means, variances and empirical distributions of individual covariates between the matched treatment and control groups [5], and through the use of various summary measures that assess the overall balance of the entire set of covariates included in the propensity score model [2,4].

It should perhaps be mentioned that there are analytical methods that use estimated propensity scores to reduce confounding that do not involve matching; for example, stratification on estimated propensity score followed by direct adjustment [3,6], regression adjustment with estimated propensity score as a predictor [3] and weighted analysis with inverse estimated propensity score as weight [2]. However, the focus here is on matching on estimated propensity score, because this approach provides the clearest analogy to constructing treatment and control groups by random assignment. Moreover, the theoretical basis for using estimated propensity scores is the same whether they are used for matching or for the above mentioned analytical methods; thus, from the perspective taken here, no generality is lost by focusing on matching.

Any covariate that is associated with both treatment choice and an outcome of interest will bias estimation of the effect of treatment on that outcome, unless the distribution of the covariate is well-balanced between treatment and control groups [2,4]. Thus, it is critical to include as many of such potentially confounding covariates as possible in the propensity score model. If during study design, investigators assemble a comprehensive inventory of known and suspected potentially confounding covariates, examination of how well this inventory is covered by the chosen dataset yields an assessment of the extent of bias reduction that is possible by matching on estimated propensity scores. This assessment can be rendered before fitting any model or examining any outcome [1,2].

In a randomized trial, the covariates displayed in 'Table 1' are generally those that were known or suspected *a priori* to be associated with a study outcome; their measurement is planned during study design, and their display serves to alert the reader that if any of these covariates are out of balance (by chance), they will probably confound estimation of the comparative treatment effect. Observational studies can be reported with similar design transparency, facilitated by the development of a propensity score model [1,2], whereby investigators describe the preparatory inventory of covariates known or suspected to be associated with treatment choice and study outcomes and note which of these covariates were available for propensity score modeling and which were not. If factors affecting the treatment choice are poorly understood, this will be revealed by an inability to compile a convincing inventory of potentially confounding covariates and will suggest that a reliable plan for bias control cannot proceed until investigators have developed a clearer understanding of how the treatment choice is made in practice.

These considerations will be illustrated by examining the designs of three recently published comparative effectiveness studies in which the primary outcome analysis was based on a subsample constructed by matching treatment and control groups 1:1 based on estimated propensity scores.

Study examples

■ Intraoperative repair of incidentally discovered patent foramen ovale

Krasuski and colleagues' cohort study of the effects of intraoperative repair of incidentally discovered patent foramen ovale (PFO) on perioperative mortality, postoperative stroke

and long-term survival provides an example of a study for which the chosen dataset provided good coverage of potentially confounding covariates [7]. When PFO, an apparent risk factor for paradoxical embolic stroke, is discovered by transesophageal echocardiography during cardiothoracic surgery, the surgeon must decide whether to alter the operative plan in order to repair the PFO [8]. Factors known to affect this decision were summarized in the introduction to this study, based on reviews of expert opinion and a survey of surgical practice. In general, the decision involved balancing the potential to reduce the risk of paradoxical stroke by repairing the PFO against the potential to increase operative risk by expanding the scope of the procedure [7–9]. Thus, factors favoring a decision to repair the PFO included, for example, a history of stroke or transient ischemic attack, risk factors for venous or atrial thrombosis, or an operative plan that already included cardiopulmonary bypass and atriotomy, in order to repair or replace the mitral or tricuspid valve.

The primary database used for this study was a clinical registry that contained detailed information abstracted from patients' operative and hospital records [7]. Long-term survival was assessed via linkage to the Social Security Death Index. The eligible sample included patients in whom a PFO was newly discovered by transesophageal echocardiography during cardiothoracic surgery between 1995 and 2006 at the Cleveland Clinic (OH, USA), a large academic medical center. The primary analysis was based on a subsample of 603 patients in whom PFOs were repaired and an equal number in whom PFOs were not repaired, matched 1:1 on estimated propensity scores. Because factors affecting surgeons' decisions were well understood and the chosen dataset included detailed clinical information from operative and hospital records, the investigators were well positioned to develop a propensity score model that provided good coverage of potentially confounding covariates.

Drug-eluting versus bare-metal stents for coronary stenosis

Tu and colleagues' cohort study of the effects of drug-eluting stents (DESs) versus bare-metal stents (BMSs) on target-vessel revascularization, myocardial infarction and death, in patients undergoing percutaneous coronary intervention (PCI), provides another example of a study in which the chosen dataset provided good coverage of potentially confounding covariates [10]. In November 2002, a sirolimus-eluting stent had been approved for use in Canada, and it was anticipated that a paclitaxel-eluting stent would also soon gain approval [101]. Both types of DES had recently been shown in randomized trials, to substantially reduce rates of restenosis and target-vessel revascularization over 6–9 months, when compared with BMSs [11,12]. However, neither type of DES had been shown to reduce rates of myocardial infarction or death when compared with BMSs, and the long-term safety and efficacy of DESs had yet to be definitively examined [11,12]. Furthermore, it was anticipated that the cost per DES would be approximately CAD$2500 more than the cost per similarly sized BMS [101].

Prior to approving funding for the incremental cost of DESs, the Ontario Ministry of Health and Long-Term Care (ON, Canada) asked the Cardiac Care Network of Ontario (CCN; ON, Canada), a professional organization that included representatives from each of the ten PCI centers in Ontario, Canada, to recommend guidelines for the use of DESs [101]. The CCN convened a working group, which recommended that the use of DESs be initially restricted to two populations for which DESs would probably provide the greatest absolute clinical benefit [101]:

- Patients with lesions at high risk of restenosis if treated with a BMS;
- Patients for whom restenosis would present a high risk for severe clinical consequences.

The working group specified that the first population comprised patients with treated diabetes, a long stenotic lesion (>18 mm) or a lesion in a narrow vessel (<2.75 mm) – three well-established risk factors for in-stent restenosis [13,14] – and that the second population primarily comprised patients with a lesion in the left main coronary artery or in a survival-dependent vessel [101].

In order to monitor introduction and use of DESs, the Ontario Ministry of Health and Long-Term Care required that PCI centers collect a clinical dataset for each DES or BMS placement, including specific patient, lesion and procedural variables, as a condition for receiving funding for the incremental cost of DESs [10]. Collected data were to be submitted to a PCI registry maintained by the CCN.

Required data elements included variables essential to application of the CCN guidelines: diabetes diagnosis, stent length (as a proxy for lesion length), stent diameter (as a proxy for vessel diameter) and identity of the stented vessel. These covariates would probably play a central role in the choice between a DES and a BMS in actual practice, since an interventional cardiologist or other representative from each of the province's PCI centers had participated in development of the CCN guidelines [101]. Other covariates in the PCI registry that were known to affect cardiovascular outcomes and that were potentially associated with the choice between a DES and a BMS, included, for example, Canadian Cardiovascular Society angina classification, days since myocardial infarction (if any), previous PCI, history of coronary artery bypass surgery, American College of Cardiology–American Heart Association lesion type and the number of vessels stented [10].

The PCI registry was the principal database used for the cohort study. Outcomes of death and myocardial infarction were identified via linkages to additional clinical and administrative databases [10]. The eligible sample included patients undergoing PCI in which one or more stents of only one type (either DES or BMS) were implanted, between 1 December 2003 and 31 March 2005. The primary analysis was based on a subsample of 3751 patients treated with DESs and an equal number treated with BMSs, matched 1:1 on estimated propensity scores. As the PCI registry had been designed to include variables essential to the application of CCN guidelines and variables known to be highly predictive of cardiovascular outcomes, the investigators were very well positioned to develop a propensity score model that provided good coverage of potentially confounding covariates.

■ Corticosteroid dosing for acute exacerbation of chronic obstructive pulmonary disease

A contrasting example of a study in which potentially confounding covariates do not appear to have been well covered by the chosen dataset is provided by a cohort study of the effect of initial corticosteroid dose on risk of treatment failure in patients hospitalized for acute exacerbation of chronic obstructive pulmonary disease (AECOPD) [15]. This study compared the effects of corticosteroids given orally at a low dose (equivalent to 20–80 mg of prednisone daily) versus intravenously at a high dose (equivalent to 120–800 mg of prednisone daily) during the first 2 days of hospitalization, to patients admitted to non-intensive care beds in over 400 hospitals in the USA in 2006 and 2007. Treatment failure, the composite primary outcome, was defined as the initiation of mechanical ventilation after 2 days of hospitalization, death during hospitalization, or readmission for chronic obstructive pulmonary disease within 30 days of discharge.

In the introduction to the study, the authors noted that guidelines for treating AECOPD recommended initiation of corticosteroids orally at a low dose [16,17,102], but the authors did not review clinical factors known or suspected to affect physicians' decisions to prescribe high-dose intravenous corticosteroids instead [15]. It should be mentioned that, as of 2006–2007, little guidance had been published regarding indications for using higher doses of systemic corticosteroids in AECOPD [17–20]. Nevertheless, clinical factors affecting physicians' dosing decisions were critical to consider; especially because an earlier study of admissions for AECOPD in 2001, based on the same administrative database that was to be used for the 2006–2007 cohort study, had found that 91% of patients treated with systemic corticosteroids received an intravenous regimen (dosages were not reported) [21]. When considering the apparent inconsistency between guidelines and practice, it is noteworthy that as of 2001 (and also still as of 2006–2007), the largest randomized trial demonstrating efficacy for systemic corticosteroids in AECOPD, conducted in 25 Veterans Affairs hospitals in the USA, had initiated treatment with high-dose intravenous corticosteroids [22], and no randomized trial of low- versus high-dose corticosteroids for AECOPD had been reported [18–20].

Guidelines aside, clinical common sense suggests numerous potentially confounding covariates that would have warranted consideration, including response to low-dose oral corticosteroids prior to admission, baseline spirometry or chronic obstructive pulmonary disease stage, and recognized measures of exacerbation severity such as degree of dyspnea, sputum volume and purulence, wheezing, and arterial blood gases [102]. The first of these – outpatient response to low-dose oral corticosteroids – may have been a frequent factor affecting inpatient dosing decisions, given that the aforementioned randomized trial in Veterans Affairs hospitals had found that

50% of patients hospitalized for AECOPD had used systemic corticosteroids within 30 days prior to admission [22].

Unfortunately, the database chosen for the cohort study had been derived from hospital administrative and billing records, and thus did not include outpatient medication history or clinical observations such as symptoms, signs, laboratory results or spirometric findings. Given the lack of clarity regarding factors affecting physicians' dosing decisions and the limitations of the chosen database, the investigators were poorly positioned to develop a propensity score model that provided adequate coverage of potentially confounding covariates.

Finally, it is noteworthy that the investigators found that treatment had been initiated with high-dose intravenous corticosteroids in 92% of study-eligible patients in the 2006–2007 cohort [15], a finding reminiscent of the 91% rate of intravenous corticosteroid regimens in the 2001 cohort [21]. Thus, the decision to initiate treatment with low-dose oral corticosteroids was very uncommon, and patients who were initiated at low dose may have been clinically exceptional. Understanding whether patients initiated at low dose had exceptional characteristics and, if so, what those characteristics were, was an essential prerequisite for identifying a subsample of patients initiated at a high dose who were truly comparable to the 8% initiated at low dose.

Conclusion

The aim of this article is not to focus on any particular study or to argue the merits of any particular treatment choice, but to suggest how development of a propensity score model can help guide study design. A thorough review of potentially confounding covariates – factors known or suspected to be associated with both treatment choice and a study outcome – forms the basis for an assessment of the extent of bias reduction that is possible by matching on estimated propensity scores. If factors affecting the treatment choice are poorly understood, investigators may want to conduct a preparatory survey of how the choice is made in practice. If potentially confounding covariates are poorly covered by the available dataset, it may be unwise to expect that a study based on this dataset can yield a valid estimate of comparative treatment effect, regardless of statistical technique [1,2,23]. Finally, if in the interest of full transparency, a summary of the preparatory inventory of potentially confounding covariates is included in the study report, readers will be better enabled to assess the potential accuracy and reliability of the study's findings.

Future perspective

Going forward, we can expect a steady increase in the number, size and scope of observational databases available to investigators for comparative effectiveness research; however, the quality and granularity of clinical information in these databases is likely to continue to be highly variable, from detailed clinical registries filled by clinicians at the point of care, to administrative databases filled primarily from billing records or claims. There is a risk that the increase in data availability and accessibility will encourage indiscriminate use of large databases to conduct observational studies without adequate consideration given to the suitability of a particular dataset for accurate and reliable estimation of comparative treatment effects [1,23]. Using the propensity score model as a guiding structure during study design can facilitate a thorough and transparent assessment of how well an available database covers potentially confounding covariates. If this assessment demonstrates that the available database cannot support a reliable estimate of comparative treatment effect, continuing on to perform an analysis of outcomes may only serve to muddle the existing evidence base. In such a case, investigators may decide to redirect their efforts to finding or developing an alternative dataset that provides sufficient coverage of those covariates that are known or suspected to be associated with choice of treatment and outcomes of interest.

Acknowledgements

The author wishes to thank the three anonymous reviewers for numerous insightful comments and helpful suggestions.

Financial & competing interests disclosure

The author has no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Background

- In an observational comparative effectiveness study, any covariate that is associated with both treatment choice and a study outcome is potentially confounding and will bias estimation of the effect of treatment on that outcome, unless the distribution of the covariate is well balanced between treatment and control groups.
- For such a study, the propensity score is the probability of choosing the treatment rather than the control condition given the values of observed covariates, and can be estimated for each patient by fitting a propensity score model to the full study sample.
- If the estimated propensity scores are used to match each treated patient with a fixed number of controls, the distributions of covariates that were included in the propensity score model will tend to be well balanced between treatment and control groups in the matched subsample, akin to the result achieved by random assignment; thus, it is critical to include as many potentially confounding covariates in the propensity score model as possible.
- If investigators assemble a comprehensive inventory of known and suspected potentially confounding covariates during study design, examination of how well this inventory is covered by the chosen dataset yields an assessment of the extent of bias reduction that is possible by matching on estimated propensity scores. Such an assessment can be rendered before fitting any model or examining any outcome.

Study examples

- The above considerations are illustrated by examining the designs of three recently published comparative effectiveness studies in which the primary outcome analysis was based on a subsample constructed by matching treatment and control groups based on estimated propensity scores.
- Two of the studies were designed with a clear understanding of factors affecting the treatment choice and study outcomes, and both employed datasets that provided good coverage of these potentially confounding covariates. Thus, for these studies, investigators were well positioned to develop effective propensity score models that could reliably contribute to bias reduction.
- The third study was apparently developed without a clear understanding of covariates affecting the treatment choice, and the chosen dataset did not cover numerous covariates that arguably should have been regarded as potentially confounding. Thus, for this study, investigators were poorly positioned to develop an effective propensity score model that could reliably contribute to bias reduction.

Conclusion

- Development of a propensity score model can help guide observational study design by focusing investigators' attention on the importance of understanding how the treatment choice is made in practice and on identifying, *a priori*, covariates that are known or suspected to be associated with treatment choice and study outcomes.
- If, in the interest of full transparency, a summary of the preparatory inventory of potentially confounding covariates is included in the study report, readers will be better enabled to assess the potential accuracy and reliability of a study's findings.

References

Papers of special note have been highlighted as:

■■ of considerable interest

1 Rubin DB. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2(3), 808–840 (2008).

■■ **Well-reasoned presentation of the conceptual basis and general approach for designing observational studies that aim to produce a reliable and valid estimation of a causal effect.**

2 Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* 2(3–4), 169–188 (2001).

■■ **Sophisticated and thorough worked example that uses estimated propensity scores to construct matched treatment and control groups for a cohort study.**

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

■■ **Seminal paper introducing theory, estimation and applications of the propensity score.**

4 Rosenbaum PR. *Design of Observational Studies.* Springer, NY, USA (2010).

■■ **Includes a less technical presentation of propensity score theory and several examples of studies that employed matching on estimated propensity scores.**

5 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28(25), 3083–3107 (2009).

6 Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79(387), 516–524 (1984).

7 Krasuski RA, Hart SA, Allen D *et al.* Prevalence and repair of intraoperatively diagnosed patent foramen ovale and association with perioperative outcomes and long-term survival. *JAMA* 302(3), 290–297 (2009).

8 Sukernik MR, Bennett-Guerrero E. The incidental finding of a patent foramen ovale during cardiac surgery: should it always be repaired? A core review. *Anesth. Analg.* 105(3), 602–610 (2007).

9 Sukernik MR, Goswami S, Frumento RJ, Oz MC, Bennett-Guerrero E. National survey regarding the management of an intraoperatively diagnosed patent foramen ovale during coronary artery bypass graft surgery. *J. Cardiothorac. Vasc. Anesth.* 19(2), 150–154 (2005).

10 Tu JV, Bowen J, Chiu M *et al.* Effectiveness and safety of drug-eluting stents in Ontario. *N. Engl. J. Med.* 357(14), 1393–1402 (2007).

11 Lemos PA, Serruys PW, Sousa JE. Drug-eluting stents: cost versus clinical benefit. *Circulation* 107(24), 3003–3007 (2003).

12 O'Neill WW, Leon MB. Drug-eluting stents: costs versus clinical benefit. *Circulation* 107(24), 3008–3011 (2003).

13 Hoffmann R, Mintz GS. Coronary in-stent restenosis – predictors, treatment and

prevention. *Eur. Heart J.* 21(21), 1739–1749 (2000).

14 Mercado N, Boersma E, Wijns W *et al.* Clinical and quantitative coronary angiographic predictors of coronary restenosis: a comparative analysis from the balloon-to-stent era. *J. Am. Coll. Cardiol.* 38(3), 645–652 (2001).

15 Lindenauer PK, Pekow PS, Lahti MC, Lee Y, Benjamin EM, Rothberg MB. Association of corticosteroid dose and route of administration with risk of treatment failure in acute exacerbation of chronic obstructive pulmonary disease. *JAMA* 303(23), 2359–2367 (2010).

16 Celli BR, MacNee W; ATS/ERS Task Force. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *Eur. Respir. J.* 23(6), 932–946 (2004).

17 National Collaborating Centre for Chronic Conditions. Chronic obstructive pulmonary disease: national clinical guideline on management of chronic obstructive pulmonary disease in adults in primary and secondary care. *Thorax* 59(Suppl. 1), 1–232 (2004).

18 Vondracek SF, Hemstreet BA. Is there an optimal corticosteroid regimen for the management of an acute exacerbation of chronic obstructive pulmonary disease? *Pharmacotherapy* 26(4), 522–532 (2006).

19 Singh JM, Palda VA, Stanbrook MB, Chapman KR. Corticosteroid therapy for patients with acute exacerbations of chronic obstructive pulmonary disease: a systematic review. *Arch. Intern. Med.* 162(22), 2527–2536 (2002).

20 Bach PB, Brown C, Gelfand SE, McCrory DC. Management of acute exacerbations of chronic obstructive pulmonary disease: a summary and appraisal of published evidence. *Ann. Intern. Med.* 134(7), 600–620 (2001).

21 Lindenauer PK, Pekow P, Gao S, Crawford AS, Gutierrez B, Benjamin EM. Quality of care for patients hospitalized for acute exacerbations of chronic obstructive pulmonary disease. *Ann. Intern. Med.* 144(12), 894–903 (2006).

22 Niewoehner DE, Erbland ML, Deupree RH *et al.* Effect of systemic glucocorticoids on exacerbations of chronic obstructive pulmonary disease. *N. Engl. J. Med.* 340(25), 1941–1947 (1999).

23 Rubin DB. On the limitations of comparative effectiveness research. *Stat. Med.* 29(19), 1991–1995 (2010).

Websites

101 Working Group on Drug Eluting Stents. *Report on initial utilization strategy: final report and recommendations.* Cardiac Care Network of Ontario, Toronto, ON, Canada (2002). www.ccn.on.ca/pdfs%5CFinal DrugElutingMaster2_Dec2002.pdf (Accessed 15 June 2011)

102 Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management and Prevention of COPD (2010). www.goldcopd.org (Accessed 16 February 2011)