PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2008; **17**: 1218–1225
Published online 29 October 2008 in Wiley InterScience (www.interscience.wiley.com) **DOI**: 10.1002/pds.1674

ORIGINAL REPORT

Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score†

Peter C. Austin PhD[1,2,3]*

[1]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[2]*Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*
[3]*Department of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*

SUMMARY

The propensity score is defined to be a subject's probability of treatment selection, conditional on observed baseline covariates. Conditional on the propensity score, treated and untreated subjects have similar distributions of observed baseline covariates. Propensity-score matching is a commonly used propensity score method for estimating the effects of treatment on outcomes. Balance diagnostics have been previously described for use when 1:1 matching on the propensity score is employed. We illustrate that these methods can be misleading when many-to-one matching on the propensity score is employed. We then propose modifications of these methods that involve weighting each untreated subject by the inverse of the number of untreated subjects in the matched set. We describe both quantitative and qualitative methods to assess the balance in baseline covariates between treated and untreated subjects in a sample obtained by many-to-one matching on the propensity score. The quantitative method uses the weighted standardized difference. The qualitative methods employ graphical methods to compare the distribution of continuous baseline covariates between treated and untreated subjects in the weighted sample. We illustrate our methods using a large sample of patients discharged from hospital with a diagnosis of a heart attack (acute myocardial infarction). The exposure was receipt of a prescription for a statin at hospital discharge. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS — balance; goodness-of-fit; observational study; propensity score; matching; propensity-score matching; pharmacoepidemiology

Received 17 June 2008; Revised 15 September 2008; Accepted 22 September 2008

INTRODUCTION

Researchers are increasingly using observational studies to estimate the effects of treatments and exposures on health outcomes. In randomized controlled trials, randomization ensures that, asymptotically (as the sample size becomes increasingly large), treated subjects will not differ systematically from untreated subjects in both measured and unmeasured baseline characteristics. Non-randomized studies of the effect of treatment on outcomes can be subject to treatment-selection bias in which treated subjects differ systematically from untreated subjects.

Propensity score methods are being used with increasing frequency to estimate treatment effects using observational data. The propensity score is defined as the probability of treatment assignment conditional on measured baseline covariates.[1–2]. Rosenbaum and Rubin demonstrated a key property of the propensity score: conditional on the true propensity score, treatment status is independent of measured baseline covariates.[1] In other words, treated and untreated subjects with the same true propensity

* Correspondence to: Dr P. C. Austin, Institute for Clinical Evaluative Sciences, G1 06, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada. E-mail: peter.austin@ices.on.ca

†The opinions, results and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred.

score will have similar distributions of observed baseline covariates.

Matching on the propensity score is a frequently employed analytic method in medical research.[3,4] Matching on the propensity score entails creating matched sets of treated and untreated subjects such that matched subjects have similar values of the propensity score. The most common implementation of propensity-score matching is 1:1 matching in which pairs of treated and untreated subjects are formed.[3] Some applied researchers are uncomfortable with this approach for a number of reasons. First, some researchers think that statistical power is reduced by discarding untreated subjects. Second, some researchers question the generalizability of the results when a large proportion of the untreated subjects are discarded. In particular, a substantial reduction in sample sizes will occur if there are substantially more untreated subjects than treated subjects. To address this concern, some researchers have employed many-to-one matching.[5–15] Using this approach, attempts are made to match multiple untreated subjects to each treated subject. Each matched set thus consists of one treated subject and multiple untreated subjects. This is similar to the approach used in many case–control studies, in which researchers match multiple controls to each case in order to increase statistical power. The use of many-to-one matching in propensity-matched analyses may be motivated by the increased statistical power that is achieved when many-to-one matching is employed in case–control studies.

When employing propensity-score matching it is important to assess whether matching on the propensity score has resulted in a matched sample in which there are no systematic differences in observed baseline characteristics between treated and untreated subjects. Methods have been described elsewhere for assessing the balance in measured baseline characteristics between treated and untreated subjects when 1:1 matching on the propensity score is employed.[16,17] However, these methods may not be appropriate when many-to-one matching is employed.

The objective of the current paper is to describe modified goodness-of-fit diagnostics for the propensity score model in the context of many-to-one matching on the propensity score. The paper is structured as follows. In Section 'Goodness-of-fit Diagnostics for the Propensity Score Model', we demonstrate why standard methods for assessing balance in baseline characteristics between treated and untreated subjects may not be appropriate in the context of many-to-one matching. We then describe modifications of these methods for use in this context. In Section 'Case Study', we describe a case study illustrating the application of these methods. Finally, in Section 'Discussion', we summarize our findings.

GOODNESS-OF-FIT DIAGNOSTICS FOR THE PROPENSITY SCORE MODEL

In this section, we first describe methods for assessing baseline balance when 1:1 matching on the propensity score was used to form a matched sample. We then illustrate how conventional balance diagnostics may be misleading when many-to-one matching is employed. Finally, we describe modifications of these diagnostics for the context of many-to-one matching. We describe both quantitative and qualitative methods for assessing balance in observed baseline covariates between treated and untreated subjects in a sample obtained using many-to-one matching on the propensity score.

Balance diagnostics for 1:1 matching on the propensity-score

Several authors have proposed that standardized differences be used to compare the mean of an observed baseline covariate between treated and untreated subjects in a propensity-score matched sample.[16,18,19] The standardized difference is defined as

$$d = \frac{\left(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}}\right)}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}} \tag{1}$$

for continuous variables, and as

$$d = \frac{\left(\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}}\right)}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}} \tag{2}$$

for dichotomous variables. The standardized difference compares the difference in means in units of the pooled-standard deviation.[20] Unlike *t*-tests and other statistical tests of hypothesis, the standardized difference is not influenced by sample size. Thus, the use of the standard difference can be used to compare balance in measured variables between treated and untreated subjects in the unweighted sample with that in the weighted sample. Furthermore, it allows for the comparison of the relative balance of variables measured in different units (e.g. age in years with systolic blood pressure in mm Hg).

While the standardized difference allows one to compare the mean of a variable between treated and untreated subjects, researchers may want to compare

DOI: 10.1002/pds

the distribution of a continuous variable between treated and untreated subjects in the matched sample. To accomplish this, researchers can use side-by-side boxplots,[21] empirical cumulative distribution functions[22] or non-parametric estimates of the probability density function. While standardized differences compare the difference in means between treated and untreated subjects, these graphical methods permit a broader comparison of the distribution of a continuous variable between two groups.

Problems with conventional balance diagnostics

Many researchers use *k*:1 matching in which one attempts to match *k* untreated subjects to each treated subjects. However, in applied applications it may not be feasible to find *k* untreated subjects for each treated subject.[5,10] For instance, one may be able to identify *k* untreated subjects for a proportion of the treated subjects. However, for the remainder of the treated subjects, fewer than *k* matched untreated subjects are located. We refer to this as incomplete *k*:1 matching.

We use a simple example to illustrate problems that can occur when conventional balance diagnostics are used to in samples obtained using incomplete *k*:1 matching. Let us assume that we have 100 treated subjects, and the value of a covariate *X* takes on the values 1, 2, 3, …, 100 for these 100 subjects ($X_i = i$, for $i = 1, \ldots, 100$). Furthermore, assume that we attempted 2:1 matching, in which we attempted to match two untreated subjects to each treated subject. However, while two matched untreated subjects were found for the first 50 treated subjects, only one matched untreated subject was found for the last 50 treated subjects. Finally, let us assume that we have perfect matching on *X* within matched sets. Thus, the value of *X* within the first 50 matched sets is (i, i, i), for $i = 1, \ldots, 50$, while the value of *X* within the last 50 matched sets is (i, i), for $i = 51, \ldots, 100$. Then, in the matched sample, the sample mean of *X* is 50.5 and 42.2 in the treated and untreated subjects, respectively. The standardized difference comparing *X* between treated and untreated subjects in the matched sample is 0.29. The empirical cumulative distribution function of *X* in treated and untreated subjects is displayed in the left panel of Figure 1. Comparing means, standardized differences and the cumulative distribution function results in the conclusion that the distribution of *X* is different between treated and untreated subjects in the matched sample. This is in the face of the perfect within-set balance on *X* between treated and untreated subjects.

Modifications of balance diagnostics for many-to-one matching on the propensity score

The sample means, sample standard deviations and sample prevalences in formulas (1) and (2) are unweighted estimates. We propose to replace each of these estimates by its weighted equivalent. For each treated subject, we define the weight to be equal to one. For each untreated subject, we define the weight to be equal to the reciprocal of the number of untreated subjects within a given matched set. Thus, in the context of 1:1 matching, all the weights would be equal to one for all subjects. However, if a matched set had one treated subject and two untreated subjects, then the weight for the untreated subjects would be 1/2, while the weight for the treated subject would be 1. Note that this does not require that there be the same number of untreated subjects within each matched set. For instance, if one matched set contained two untreated subjects, while a second matched set contained three untreated subjects, then the weights for the untreated subjects in these two matched sets would be 1/2 and 1/3, respectively.

The weighted mean is defined as $\bar{x}_w = \sum w_i x_i / \sum w_i$, while the weighted sample variance is defined as $s_w^2 = (\sum w_i / ((\sum w_i)^2 - \sum w_i^2)) \sum w_i (x_i - \bar{x}_w)^2$, where w_i denotes the weight for the *i*th subject.

The weighted standardized difference can be used to quantitatively compare the balance in baseline characteristics between treated and untreated subjects in the matched sample. Weighted estimates of side-by-side boxplots and empirical cumulative distribution functions can be used to qualitatively compare the distribution of continuous baseline covariates between treated and untreated subjects in the matched sample.

In our simple example described in Section ‘Problems with Conventional Balance Diagnostics’, the weighted standardized difference for *X* is 0. The weighted empirical cumulative distribution function comparing the distribution of *X* between treated and untreated subjects in the matched sample is depicted in the right panel of Figure 1. Thus, one can observe that by accounting for the weights, our balance diagnostics demonstrate that there is no difference in *X* between treated and untreated subjects.

CASE STUDY

Data sources

We used data on 9104 patients who were discharged alive with an acute myocardial infarction (AMI or heart attack) from 102 hospitals in Ontario, Canada,

DOI: 10.1002/pds

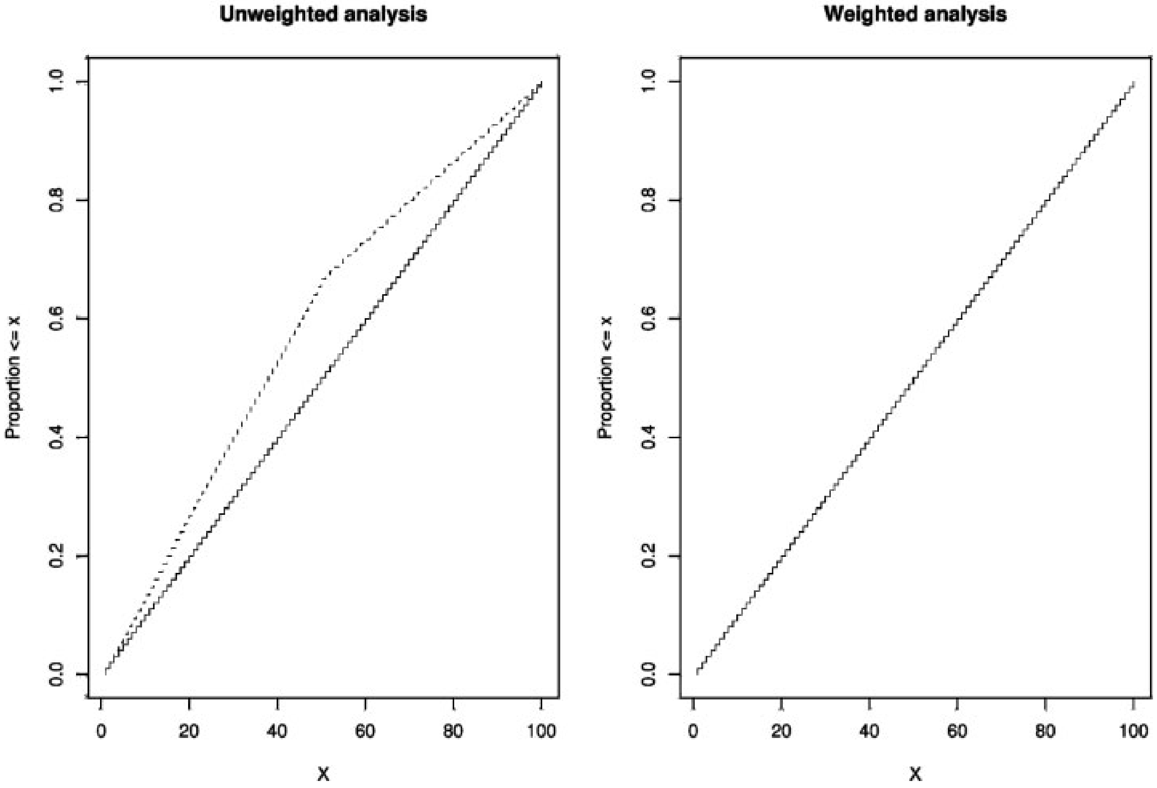

Figure 1. Unweighted and weighted empirical cumulative distribution functions

between 1 April 1999 and 31 March 2001. These data are similar to those reported on elsewhere,[23–25] and were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study, an initiative that is focused on improving the quality of care for cardiovascular disease patients in Ontario.[26] Data on patient demographics, presenting signs and symptoms, classic cardiac risk factors, comorbid conditions and vascular history, vital signs on admission and results of laboratory tests were abstracted directly from patients' medical records. The exposure of interest was whether the patient was prescribed a statin at hospital discharge. Research ethics approval for this study was obtained from Sunnybrook Health Sciences Centre.

Overall, 3049 (33.5%) of patients received a prescription for a statin at discharge, while 6055 (66.5%) did not receive a prescription at discharge. Characteristics of patients who did and did not receive a statin are described elsewhere.[16] Briefly, patients receiving a statin prescription at discharge tended to be younger and healthier than those who did not receive a statin prescription at discharge. Standardized differences for the 24 baseline covariates are reported in the second column of Table 1. Eight of the 24 measured baseline covariates had standardized differences that exceeded 10%, indicative of imbalance in these covariates between treated and untreated subjects.[18,19]

Methods

A propensity score model was fit using a logistic regression model in which treatment assignment (statin *vs*. no statin) was regressed on the 24 baseline covariates listed in Table 1. Each covariate entered the propensity score model as a main effect only. The continuous variables were assumed to be linearly related to the log-odds of receiving a prescription for a statin at hospital discharge. We then randomly selected 49 of the treated patients and excluded them from subsequent analyses. Thus, there were 3000 treated subjects and 6055 untreated subjects. This exclusion was made so that there would be at least two untreated subjects for each treated subject in the sample.

DOI: 10.1002/pds

Table 1. Standardized differences for assessing balance of baseline covariates in initial and matched samples

Variable	Initial sample—unmatched	1:1 matching	2:1 matching—unweighted	2:1 matching—weighted
Demographic characteristics				
Age	0.361	0.009	0.006	0.013
Female	0.169	0.016	0.029	0.010
Presenting signs and symptoms				
Cardiogenic shock	0.048	0.017	0.001	0.000
Acute CHF/pulmonary edema	0.058	0.002	0.019	0.012
Cardiac risk factors				
Family history of coronary artery disease	0.202	0.009	0.025	0.012
Diabetes	0.009	0.003	0.036	0.004
Hyperlipidemia	0.875	0.005	0.343	0.008
Hypertension	0.068	0.007	0.081	0.006
Current smoker	0.042	0.009	0.021	0.017
Comorbid conditions and co-existing illnesses				
CVA/TIA	0.081	0.005	0.026	0.015
Angina	0.101	0.018	0.022	0.005
Cancer	0.046	0.008	0.002	0.005
Chronic CHF	0.082	0.017	0.004	0.016
Renal disease	0.019	0.006	0.003	0.009
Vital signs on admission				
Heart rate	0.122	0.010	0.000	0.008
Systolic blood pressure	0.021	0.011	0.020	0.012
Diastolic blood pressure	0.048	0.021	0.012	0.000
Respiratory rate	0.172	0.015	0.002	0.001
Laboratory tests on admission				
White blood count	0.067	0.020	0.022	0.020
Hemoglobin	0.171	0.006	0.040	0.020
Sodium	0.081	0.020	0.017	0.015
Glucose	0.037	0.014	0.020	0.008
Potassium	0.062	0.025	0.025	0.013
Creatinine	0.100	0.004	0.024	0.004

Two different matched samples were constructed. First, we used 1:1 matching. Treated and untreated subjects were matched on the logit of the propensity score using calipers of width equal to 0.2 times the standard deviation of the logit of the propensity score. Second, we used 2:1 matching in which we attempted to match two untreated subjects to each treated subject. As above, treated and untreated subjects were matched on the logit of the propensity score using calipers of width equal to 0.2 times the standard deviation of the logit of the propensity score.

In each matched sample, we used standardized differences to compare the balance in baseline covariates between treated and untreated subjects. In the sample obtained using 2:1 matching, both weighted and unweighted standardized differences were used.

Results

Two thousand four hundred ten matched pairs were formed when 1:1 matching was employed. Standardized differences comparing the balance in baseline covariates in this matched sample are reported in the third column of Table 1. The largest standardized difference was 0.002.

When 2:1 matching was used, 2410 matched sets were also formed. Of these matched sets, 1281 consisted of two untreated subjects and one treated subject, while 1129 sets consisted of one untreated subject and one treated subject. Thus, a total of 6101 subjects were included in the matched sample. However, in both matching schemes, 80.3% of treated subjects were matched to at least one untreated subject. The unweighted and weighted standardized differences for each of the measured baseline covariates are reported in the fourth and fifth columns of Table 1, respectively. There were a few notable discrepancies between the unweighted and weighted standardized differences in the 2:1 matched sample. The largest absolute discrepancy was for history of hyperlipidemia. The unweighted standardized difference was 0.343, while the weighted standardized difference was 0.008, which was closer to that observed in the 1:1 matched sample (0.005). In the unmatched sample, the standardized difference for history of hyperlipidemia was 0.875. Similarly, the unweighted standardized difference for history of hypertension was 0.081 in the 2:1 matched sample, while the weighted standardized difference was 0.006. In the overall unmatched sample, the standardized difference for hypertension was 0.068. Thus, the use of the unweighted standardized difference in the 2:1 matched sample would indicate that balance was *worse* in the 2:1 matched sample than it was in the initial unmatched sample.

Figure 2 depicts both weighted and unweighted empirical cumulative distribution functions for two of the continuous variables: potassium and haemoglobin. The unweighted distribution functions are depicted in the two panels on the left, while the weighted distribution functions are depicted in the two panels on the right. For each of the two variables, the unweighted and weighted cumulative distribution functions are essentially identical. Similar graphics were obtained for the other continuous variables.

DISCUSSION

In the current paper, we have modified balance diagnostics for use in many-to-one matching on the propensity score. We have demonstrated that ignoring incomplete matching can result in misleading balance diagnostics. Weighting each matched untreated subject by the inverse of the number of untreated subjects in the matched set allows one to produce correct balance diagnostics.

Several studies in the medical literature have employed many-to-one matching on the propensity score.[5–15] Of these, some have used incomplete many-to-one matching.[5,10] Of these studies, none accounted for incomplete matches when assessing balance between treated and untreated subjects in the matched sample. The methods proposed in the current paper will allow for better balance diagnostics in studies that employ incomplete many-to-one matching.

Rubin has argued that an advantage to the use of propensity-score methods is that one can design an observational study without the outcome being in sight.[27] The diagnostics that we have developed are consistent with that paradigm. None of the diagnostics that we present refer to an outcome variable. Indeed, in the case study, the only variables referenced were the exposure variable (prescription for a statin at discharge) and measured baseline covariates. Rubin

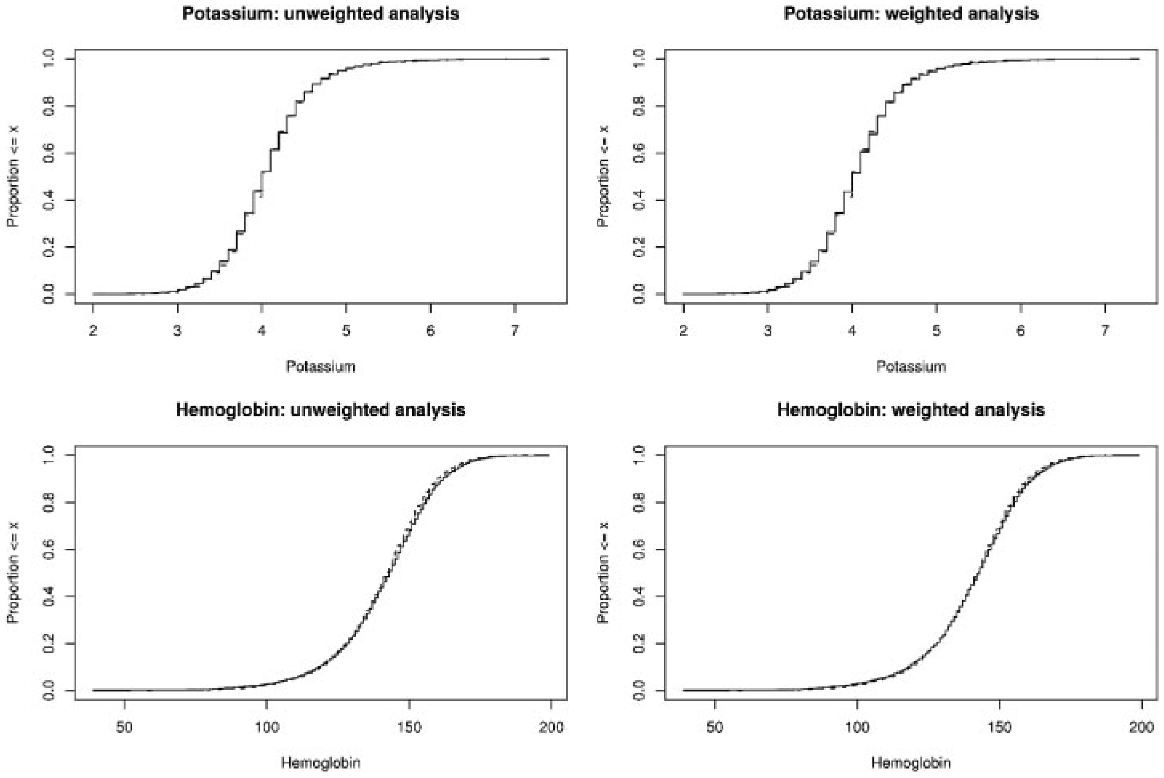

Figure 2. Empirical cumulative distribution functions

DOI: 10.1002/pds

suggests '*diagnostics for the successful design of observational studies based on estimated propensity scores ... is a critically important activity in most observational studies*'.[28] Matching on the propensity score results in a matched sample in which treated and untreated subjects have the same distribution of observed baseline covariates. Balance diagnostics serve an important role in assessing whether the propensity score model has been correctly specified.[29] Note that matching in the correctly specified propensity score will only balance the distribution of measured baseline covariates between treated and untreated subjects. It need not result in balancing unmeasured variables between treated and untreated subjects.[19,30]

The balance diagnostics that we have proposed have been based on properties of the matched sample and not on statistical hypothesis testing. Other authors have criticized the use of balance diagnostics that are based on hypothesis testing.[17] Hypothesis testing and p-values are confounded with sample size. The matched sample is almost invariably smaller than the initial sample. Thus, when using hypothesis testing, the appearance of improvements in balance may be due only to the diminished sample size.[17] For this reason we have not proposed balance tests based on significance testing. Instead, we have focused on methods such as standardized differences and empirical cumulative distribution functions that are properties of samples and are not influenced by sample size. While balance diagnostics have been proposed for use with stratification on the propensity score[2,19,30] and 1:1 matching on the propensity score,[16,17] we are not aware of balance diagnostics that have been described for use with many-to-one matching on the propensity score.

The balance diagnostics described in this paper are restricted for use in studies that employ matching on the propensity score. An alternative method for employing propensity scores is subclassification or stratification on the propensity score.[2,16] In this approach, the effect of treatment is estimated in each of the subclasses or strata, and the subclass-specific treatment effects are pooled.[2] The balance diagnostics described in this paper are not appropriate for subclassification on the propensity score. However, alternate diagnostics have been described elsewhere.[2,16,30] In brief, these methods are based on comparing the similarity of treated and untreated subjects within each quintile, rather than across the entire sample.

The focus of the current paper has been on balance diagnostics when many-to-one matching on the propensity score is employed. We have not directly examined the question of how propensity-score matched sets are formed. In two systematic reviews of the literature, it was found that a wide range of calipers were used when propensity-score matching was employed in the medical literature.[3–4] A recent study compared the relative performance of the more commonly used methods for propensity-score matching.[31] In an empirical examination, the use of seven of the eight propensity-score matching methods examined resulted in qualitatively similar estimates of treatment effect. The eighth propensity-score matching method resulted in a qualitatively different estimate of treatment effect compared to the other seven methods. In subsequent Monte Carlo simulations, it was found that matching using calipers of width of 0.2 of the standard deviation of the logit of the propensity score and the use of calipers of width 0.02 and 0.03 tended to have superior performance for estimating treatment effects.[31] However, further research is required into the performance of different methods for propensity-score matching.

In summary, we have described diagnostics for assessing whether the propensity score model has been adequately specified when using many-to-one matching on the propensity score. These methods allow investigators to assess whether the propensity score has been adequately specified and whether matching on the propensity score has resulted in a matched sample in which systematic differences between treated and untreated subjects have been reduced or eliminated.

KEY POINTS

- Diagnostics for whether the propensity score model has been correctly specified are based on comparing whether the distribution of measured baseline covariates is similar between treated and untreated subject with similar values of the propensity score.
- In the context of matching on the propensity score, the distribution of baseline characteristics is compared between treated and untreated subjects in the matched sample.
- When many-to-one matching is used, one must account for the number of untreated subjects matched to each treated subjects.
- Sample-specific measures of balance should incorporate weights that are based on the number of treated and untreated subjects within each matched set.

ACKNOWLEDGEMENTS

The Institute for Clinical Evaluative Sciences (ICES) is supported, in part, by a grant from the Ontario Ministry of Health and Long Term Care. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). The data used in the study was provided by the EFFECT Study which was funded by a CIHR Team Grant in Cardiovascular Outcomes Research. Dr Austin is supported, in part, by a Career Scientist award from the Heart and Stroke Foundation of Ontario.

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. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**: 516–524.
3. Austin PC. A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Stat Med* 2008; **27**: 2037–2049.
4. 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.
5. Boening A, Friedrich C, Hedderich J, Schoettler J, Fraund S, Cremer JT. Early and medium-term results after on-pump and offpump coronary artery surgery: a propensity score analysis. *Ann Thorac Surg* 2003; **76**: 2000–2006.
6. Aronow HD, Novaro GM, Lauer MS, *et al*. In-hospital initiation of lipid-lowering therapy after coronary intervention as a predictor of long-term utilization: a propensity analysis. *Arch Intern Med* 2003; **163**: 2576–2582.
7. Magee MJ, Jablonski KA, Stamou SC, *et al*. Elimination of cardiopulmonary bypass improves early survival for multivessel coronary artery bypass patients. *Ann Thorac Surg* 2002; **73**: 1196–1202.
8. Sernyak MJ, Desai R, Stolar M, Rosenheck R. Impact of clozapine on completed suicide. *Am J Psychiatry* 2001; **158**: 931–937.
9. Chukwuemeka A, Weisel A, Maganti M, *et al*. Renal dysfunction in high-risk patients after on-pump and off-pump coronary artery bypass surgery: a propensity score analysis. *Ann Thorac Surg* 2005; **80**(6): 2148–2153.
10. Reeves BC, Ascione R, Caputo M, Angelini GD. Morbidity and mortality following acute conversion from off-pump to on-pump coronary surgery. *Eur J Cardiothorac Surg* 2006; **29**(6): 941–947.
11. Rajakaruna C, Rogers CA, Angelini GD, Ascione R. Risk factors for and economic implications of prolonged ventilation after cardiac surgery. *J Thorac Cardiovasc Surg* 2005; **130**(5): 1270–1277.
12. Kaw R, Golish J, Ghamande S, Burgess R, Foldvary N, Walker E. Incremental risk of obstructive sleep apnea on cardiac surgical outcomes. *J Cardiovasc Surg* 2006; **47**: 683–689.
13. Ahmed A, Perry GJ, Fleg JL, Love TE, Goff DC Jr, Kitzman DW. Outcomes in ambulatory chronic systolic and diastolic heart failure: a propensity score analysis. *Am Heart J* 2006; **152**(5): 956–966.
14. Stamou SC, White T, Barnett S, Boyce SW, Corso PJ, Lefrak EA. Comparisons of cardiac surgery outcomes in Jehovah's versus Non-Jehovah's witnesses. *Am J Cardiology* 2006; **98**: 1223–1225.
15. Toumpoulis IK, Anagnostopoulos CE, Katritsis DG, DeRose JJ Jr, Swistel DG. The impact of preoperative thrombolysis on long-term survival after coronary artery bypass grafting. *Circulation* 2005; **112** [9 Suppl I] I-351–I-357.
16. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006; **25**: 2084–2106.
17. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc, Ser A (Stat Soc)* 2008; **171**: 481–502.
18. Normand SLT, Landrum MB, Guadagnoli E, *et al*. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol* 2001; **54**: 387–398.
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**: 734–753.
20. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat* 1986; **40**: 249–251.
21. Hoaglin DC, Mosteller F, Tukey JW. *Understanding Robust and Exploratory Data Analysis*. John Wiley & Sons: New York, NY, 1983.
22. Casella G, Berger RL. *Statistical Inference*. Duxbury Press: Belmont, CA, 1990.
23. Austin PC, Mamdani MM, Juurlink DN, Alter DA, Tu JV. Missed opportunities in the secondary prevention of myocardial infarction: an assessment of the effects of statin underprescribing on mortality. *Am Heart J* 2006; **151**: 969–975.
24. Austin PC, Tu JV. Comparing clinical data with administrative data for producing AMI report cards. *J R Stat Soc, Ser A (Stat Soc)* 2006; **169**: 115–126.
25. Austin PC. A comparison of classification and regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med* 2007; **26**: 2937–2957.
26. Tu JV, Donovan LR, Lee DS, *et al*. *Quality of Cardiac Care in Ontario*. Institute for Clinical Evaluative Sciences: Toronto, Ontario, 2004.
27. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol* 2001; **2**: 169–188.
28. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf* 2004; **13**: 855–857.
29. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal* 2007; **15**: 199–236.
30. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. *Stat Med* 2005; **24**: 1563–1578.
31. Austin PC. The performance of different propensity-score matching methods used in the medical literature. *Biom J* In-press.

DOI: 10.1002/pds