PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2008; **17**: 1202–1217
Published online 29 October 2008 in Wiley InterScience (www.interscience.wiley.com) **DOI**: 10.1002/pds.1673

ORIGINAL REPORT

Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score†

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

[1]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*
[2]*Department of Public Health Sciences, 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. In the medical literature, there are three commonly employed propensity-score methods: stratification (sub-classification) on the propensity score, matching on the propensity score, and covariate adjustment using the propensity score. Methods have been developed to assess the adequacy of the propensity score model in the context of stratification on the propensity score and propensity-score matching. However, no comparable methods have been developed for covariate adjustment using the propensity score. Inferences about treatment effect made using propensity-score methods are only valid if, conditional on the propensity score, treated and untreated subjects have similar distributions of baseline covariates. We develop both quantitative and qualitative methods to assess the balance in baseline covariates between treated and untreated subjects. The quantitative method employs the weighted conditional standardized difference. This is the conditional difference in the mean of a covariate between treated and untreated subjects, in units of the pooled standard deviation, integrated over the distribution of the propensity score. The qualitative method employs quantile regression models to determine whether, conditional on the propensity score, treated and untreated subjects have similar distributions of continuous covariates. We illustrate our methods using a large dataset of patients discharged from hospital with a diagnosis of a heart attack (acute myocardial infarction). The exposure was receipt of a prescription for a beta-blocker at hospital discharge. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS — balance; covariate adjustment; goodness-of-fit; observational study; propensity score; quantile regression

Received 24 July 2008; Revised 19 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 studies of treatment effect, randomization ensures that, on average, treated subjects will not differ systematically from untreated subjects in both measured and unmeasured baseline characteristics. Therefore, any differences in outcomes can be attributed to the treatment or exposure. Non-randomized studies of the effect of treatment on outcomes can be subject to treatment-selection bias because treated subjects frequently differ systematically from untreated subjects.

* Correspondence to: 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 author has no conflicts of interest to report

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 propensity score, treatment status is independent of measured baseline covariates.[1] In other words, treated and untreated subjects with the same propensity score will have similar distributions of observed baseline covariates.

Three methods of using the propensity score are commonly employed in the medical literature: covariate adjustment using the propensity score, stratification or subclassification on the propensity score, and matching on the propensity score.[3] In propensity-score matching, pairs of treated and untreated subjects with a similar propensity score are formed (while there are other variations to propensity-score matching, this is the most common approach). The effect of the treatment on the outcome is then estimated using the propensity-score matched sample. In stratification on the propensity score, a pooled estimate of the treatment effect is obtained across different strata (often the quintiles) of the propensity score. Covariate adjustment using the propensity score uses regression adjustment in which the outcome variable is regressed on the estimated propensity score and an indicator variable denoting treatment selection.[3] Each of these three methods was proposed by Rosenbaum and Rubin[1] in their original article on the propensity score. Systematic reviews of the use of propensity-score methods in the medical literature have found that covariate adjustment using the propensity score is the most commonly implemented propensity-score method in the medical literature.[4–6]

In simple randomized experiments, the true propensity score is known and is fixed by the design of the experiment. However, in observational studies, the true propensity score is unknown and must be estimated using the data. The propensity score is frequently estimated using a logistic (or probit) regression model in which treatment selection is regressed on measured baseline covariates. Since the true propensity score model is not known, the researcher must specify the nature of the propensity score model. The test of whether the propensity score model has been correctly specified is whether treated and untreated subjects with similar propensity scores have similar distributions of measured baseline covariates.[2,7]

Goodness-of-fit diagnostics for the adequacy of the propensity score model have been developed in the context of propensity-score matching and stratification on the propensity score. Methods have been developed to assess whether matching on the propensity score has resulted in a matched sample in which the distribution of measured baseline covariates are similar between treated and untreated subjects. Several studies have described the use of standardized differences to compare the distribution of baseline variables between treated and untreated subjects in the matched sample.[3,8,9] Similarly, comparison of non-parametric density functions of the propensity score in treated and untreated subjects has been proposed.[3] Ho *et al.*[7] suggests comparing higher order moments and important two-way interactions between treated and untreated subjects. Similarly, balance diagnostics have been proposed for when stratification on the quintiles of the propensity score is employed. In one of the original propensity score articles, Rosenbaum and Rubin used two-way ANOVA models to regress each measured baseline covariate on propensity score quintile (as a five-level categorical variable), an indicator variable for treatment selection, and the two-way interaction between these two factors.[2] The significance of either the treatment indicator or the interaction variable was used to denote that the distribution of that baseline covariate differed between treated and untreated subjects within at least one quintile of the propensity score. Other authors have proposed the following methods: within quintile side-by-side boxplots to compare the distribution of the propensity score between treated and untreated subjects within each quintile of the propensity score,[10] the use of within quintile standardized differences to compare the distribution of baseline covariates between treated and untreated subjects,[3,9] and within quintiles quantile–quantile plots of the estimated propensity score in treated and untreated subjects.[10] While several applied studies have reported the area under the receiver operating characteristic (ROC) curve of the propensity score model (equivalent to the model c-statistic), recent research has indicated that this does not serve as a goodness-of-fit test of the propensity score model.[9] The area under the ROC curve for the propensity score model is a measure of model discrimination. However, different propensity score models can have different ROC curve areas, yet matching on the different estimated propensity scores can result in matched samples in which prognostically important covariates have equivalent balance between treated and untreated subjects.[9] Despite the frequency with which covariate

DOI: 10.1002/pds

adjustment using the propensity score is employed in the medical literature, no goodness-of-fit diagnostics have been developed for the propensity score model in this context.

The objective of the current study is to propose goodness-of-fit diagnostics for the propensity score model in the context of covariate adjustment using the propensity score. The paper is structured as follows. In the next section we propose two goodness-of-fit diagnostics for the propensity score model. The first method extends the standardized difference to the context of covariate adjustment using the propensity score. We refer to this method as the weighted conditional standardized difference. The second method uses quantile regression to compare the distribution of continuous covariates between treated and untreated subjects with similar propensity scores. The section following it describes Monte Carlo simulations used to evaluate the performance of weighted conditional standardized difference. In the following section, we describe a case-study illustrating the application of these methods. Finally, we summarize our findings.

BALANCE DIAGNOSTICS FOR THE PROPENSITY SCORE MODEL

The balance diagnostics described above for propensity-score matching and for stratification on the propensity score share a similarity, despite employing different methods. Each diagnostic allows one to compare the distribution of measured baseline covariates between treated and untreated subjects. For propensity-score matching, this was done in the matched sample, while for stratification on the propensity score, this was done within stratum (usually the quintiles) defined by the propensity score. The principal idea behind these diagnostics was to determine whether, in subjects with a similar propensity score, treated and untreated subjects had a similar distribution of baseline covariates.

When covariate adjustment using the estimated propensity score is employed, the following regression model is fit to the sample data

$$Y = \alpha_0 + \alpha_1 T + \alpha_2 Z + \varepsilon \tag{1}$$

where Y denotes the outcome, T is an indicator variable denoting treatment selection ($T = 1$ denoting treated; $T = 0$ denoting untreated), and Z denotes the estimated propensity score (generalized linear models can be fit for dichotomous or count outcomes, while survival models can be fit for time-to-event outcomes). The above model assumes that if the propensity score model has been adequately specified, then treated and untreated subjects with the same estimated propensity score will have similar distributions of measured baseline covariates. In other words, by conditioning on the propensity score, one has eliminated systematic differences between treated and untreated subjects.

The assumption that the propensity score model has been correctly specified is critical to the application of covariate adjustment using the propensity score. However, in practice, the true propensity score is not known. Ho *et al*. describe the propensity score tautology as follows: the true propensity score is a balancing score—conditional on the propensity score—treated and untreated subjects will have the same distribution of measured baseline characteristics. While the true propensity score is not usually known, we know that the propensity score model has been correctly specified when, conditional on the propensity score, treated and untreated subjects have similar distributions of measured baseline covariates.[7] The methods developed in this Section will enable applied researchers to examine the degree to which, conditioning on the estimated propensity score has removed systematic differences between treated and untreated subjects.

Weighted conditional standardized differences

The first proposed diagnostic is a weighted version of the standardized difference. This allows one to compare the difference in means of baseline covariates between treated and untreated subjects with the same propensity score. This diagnostic can be computed for both continuous and dichotomous baseline covariates. The standardized difference is defined as

$$d = \frac{\overline{x}_{\text{treatment}} - \overline{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treatment}} + s^2_{\text{control}}}{2}}} \tag{2}$$

for continuous variables, and by

$$d = \frac{\hat{p}_{\text{treatment}} - \hat{p}_{\text{control}}}{\sqrt{\frac{\hat{p}_T(1-\hat{p}_T) + \hat{p}_C(1-\hat{p}_C)}{2}}} \tag{3}$$

for dichotomous variables. The standardized difference is the difference in means between the two groups divided by an estimate of the common standard deviation of that variable in the two groups. It represents the number of standard deviations by which the two groups differ.[11]

DOI: 10.1002/pds

Let X denote a continuous baseline covariate. The following linear regression model, estimated using ordinary least squares (OLS) can be fit to the data

$$X=\alpha_0+\alpha_1 T+\alpha_2 Z+\alpha_3 T\times Z+\varepsilon,\ \varepsilon\sim N(0,\sigma^2) \quad (4)$$

where Z denotes the estimated propensity score and T denotes treatment status ($T=1$ denotes treated; $T=0$ denotes untreated). We have included an interaction between Z, the estimated propensity score; and T, the treatment indicator; to allow the mean difference in the covariate between treated and untreated subjects to be different for different values of the propensity score. For a given value of Z, the mean predicted baseline covariate is $\hat{\alpha}_0+\hat{\alpha}_1+\hat{\alpha}_2 Z+\hat{\alpha}_3 Z$ and $\hat{\alpha}_0+\hat{\alpha}_2 Z$ for treated and untreated subjects, respectively. Thus, conditional on Z, the standardized difference comparing the mean of X for treated subjects with that for untreated subjects is $\frac{\hat{\alpha}_1+\hat{\alpha}_3 Z}{\hat{\sigma}}$. We refer to this as the conditional standardized difference. The absolute value of this term is referred to as the conditional standardized absolute difference. This quantity can then be integrated over the distribution of the estimated propensity score in the study sample to obtain a weighted conditional absolute standardized difference that reflects the average standardized absolute difference between treated and untreated subjects with the same propensity score

$$\int_Z \frac{|\hat{\alpha}_1+\hat{\alpha}_3 Z|}{\hat{\sigma}}\,dZ$$

where $_Z$ denotes the empirical distribution of the propensity score in the study sample. Note that it is important to integrate over the distribution of the propensity score, and not over the range of propensity score. This is since the distribution of the propensity score may not be uniform over its range. In some contexts, there may be relatively few subjects with a very low or very high propensity score. We propose integrating the standardized absolute difference rather than the standardized difference so that positive and negative differences do not cancel one another. For instance, assume that among subjects with low propensity score, the conditional mean is greater for treated subjects than for untreated subjects, while the converse is true for subjects with high propensity scores. Then, integrating the conditional standardized difference rather than the conditional standardized absolute difference could mask these differences between treated and untreated subjects. Furthermore, we propose that the conditional standardized absolute difference be integrated rather than the conditional absolute difference. By expressing the difference in units of standard deviation, one can compare the relative balance of variables measured in different metrics.

We now describe how the conditional standardized absolute difference can be co-computed in practice. First, for a given baseline covariate X, one fits the linear regression model described in Equation (4), in which X is regressed on an indicator variable denoting treatment status, the propensity score, and the interaction between these two terms. Second, for each subject in the sample, the following quantity is estimated: $\frac{|\hat{\alpha}_1+\hat{\alpha}_3 Z|}{\hat{\sigma}}$. In this formula, $\hat{\alpha}_1$ and $\hat{\alpha}_3$ denote the estimated regression coefficients for the treatment indicator and for the interaction between the treatment indicator and the estimated propensity score, respectively. The square root of the estimate of the residual variance (the variance of the error term) is denoted by $\hat{\sigma}$. Finally, the mean of the above quantity is determined across all subjects in the sample. This is the estimate of the weighted standardized absolute difference.

The above method can be adapted for dichotomous baseline covariates. The linear regression model in Equation (4) can be replaced by a logistic regression model given below

$$\text{logit}(\Pr(X=1))=\alpha_0+\alpha_1 T+\alpha_2 Z+\alpha_3 T\times Z \quad (5)$$

For a given value of Z, the predicted probability of the dichotomous variable can be computed for treated and untreated subjects. Conditional on Z, the standardized difference can be computed using formula (3), and this quantity can be integrated over the empirical distribution of the propensity score.

We have not proposed tests of balance based on testing the hypothesis:$H_0:\alpha_1=\alpha_3=0$. There are limitations to balance tests based on statistical tests of hypotheses. First, the power of such tests is influenced by sample size. Therefore, given the same quantitative degree of balance between treated and untreated subjects, imbalance is more likely to be detected in larger samples than in smaller samples. Second, in a given sample, the power of the above test may differ from that of a simple *t*-test comparing means between treated and untreated subjects in the overall sample. Therefore one could have a situation in which one is comparing balance before and after conditioning on the propensity score using two tests that have different statistical power. Finally, our proposed balance diagnostics are consistent with the framework of Imai *et al.*[12] who suggested that, in the context of

DOI: 10.1002/pds

propensity-score matching, balance diagnostics should be based on properties of a sample.

Quantile regression to compare conditional distributions of continuous variables

The methods described in the above section allow one to quantify the mean difference in conditional means between treated and untreated subjects. In this section, we describe methods to qualitatively compare the distribution of measured baseline covariates between treated and untreated subjects with the same propensity score.

Quantile regression is a regression model in which a specified quantile of the dependent variable is regressed on subject characteristics.[13,14] This is in contrast to OLS regression, in which the expectation of the response variable is regressed on subject characteristics. An advantage to the use of quantile regression compared to OLS regression is that one can examine how several quantiles (or percentiles) of the conditional response distribution vary with the predictor variables.[13] In OLS regression, one can only examine the effect of the predictors on the conditional mean of the dependent distribution.

Let X denote a continuous baseline covariate. We propose to use a family of quantile regression models of the form

$$q(X, \rho|T, Z) = \alpha_0 + \alpha_1 T + \alpha_2 Z + \alpha_3 T \times Z \qquad (6)$$

where $q(X, \rho)$ denotes the conditional ρth quantile of the covariate X. The quantile regression model in Equation (6) can be fit for different regression quantiles. We propose the use of the 5th, 25th, 50th (median), 75th, and 95th regression quantiles. These quantiles are frequently used in summarizing distributions of continuous variables. However, other quantiles could be selected in practice. Plotting the estimated regression quantiles against the estimated propensity score for treated and untreated subjects separately allows one to examine the distribution of X at specific values of Z, the estimated propensity score, in treated and untreated subjects.

The use of quantile regression allows one to qualitatively examine the conditional distribution of a continuous baseline covariate. For a dichotomous baseline covariate that denotes the presence or absence of a risk factor, one can fit the logistic regression model described in formula (5). The model-derived predicted probabilities of the presence of the risk factor can be determined for treated and untreated subjects at value of Z, the propensity score. The conditional probability of the presence of the risk factor can then be graphically displayed in treated and untreated subjects across the range of the propensity score.

The use of weighted conditional standardized absolute differences and quantile regression are intended to be complementary. The use of quantile regression will provide little additional information if two conditions are satisfied. First, if conditional on the propensity score, the distribution of a baseline characteristic is symmetric within each treatment group and if the conditional distribution has the same shape for both treated and untreated subjects. Second, if for a given treatment group, the conditional distribution of the baseline covariate is of the same shape for different values of the propensity score (i.e., only the location is shifted). Quantile regression provides a method to assess whether these two conditions are satisfied. The use of the weighted conditional standardized absolute difference only allows one to assess whether, conditional on the propensity score, the mean of a covariate is similar between treated and untreated subjects. Quantile regression permits a more detailed examination of the similarity of the conditional distribution between treated and untreated subjects.

Comparison with balance diagnostics for stratification on the propensity score

There are some similarities between the proposed balance diagnostics and those proposed for use with stratification (subclassification) on the propensity score. In the most frequent implementation of stratification on the propensity score, the sample is divided into five approximately sized strata using the quintiles of the estimated propensity score. Within each strata, a stratum-specific treatment effect is estimated. Thus, the effect of treatment is estimated by comparing treated and untreated subjects within the same quintile of the propensity score. The stratum-specific treatment effects are then pooled to obtain an overall treatment effect.[1,2] Balance diagnostics are based upon comparing the distribution of baseline covariates between treated and untreated subjects within the same stratum (typically the quintiles of the propensity score). Authors have proposed examining quintile-specific standardized differences and within-quintile side-by-side boxplots for comparing the distribution of variables between treated and untreated subjects.[9,10] The first approach allows for the

DOI: 10.1002/pds

comparison of means between treated and untreated subjects within the same quintile. The second approach allows for a qualitative comparison of the distribution of continuous covariates between treated and untreated subjects within the same stratum. Both balance diagnostics compare the distribution of baseline covariates between treated and untreated subjects within the same stratum. The rationale for this is that one is comparing outcomes between treated and untreated subjects in the same stratum. When using covariate adjustment using the propensity score, one is not comparing outcomes between treated and untreated subjects within the same propensity score quintile. Instead, one is comparing outcomes between treated and untreated subjects with the same value of the propensity score. Therefore, it is inadequate simply to examine whether treated and untreated subjects within the same propensity-score quintile have similar distributions of baseline variables. Instead, one must examine whether treated and untreated subjects with the *same* propensity score have similar distributions of measured baseline covariates.

If conventional balance diagnostics for stratification on the propensity score indicate that there are differences in measured baseline covariates between treated and untreated subjects within the same stratum, then it is likely that differences exist between treated and untreated subjects with the same propensity score. However, if acceptable within-stratum balance is found, then it does not necessarily follow that acceptable balance will be observed between subjects with the same propensity score. In the Appendix, we describe a setting in which acceptable within-quintile balance is observed, whereas treated and untreated subjects with the same propensity score have different distributions of a baseline covariate. An important criterion for developing balance diagnostics is that they must compare the distribution of baseline covariates in a way that reflects how the propensity score is being used.

MONTE CARLO SIMULATIONS

Methods

We used Monte Carlo simulations similar to those described in a prior study to examine the performance of different methods for propensity-score matching.[15] We randomly generated data such that there were 10 variables that were imbalanced between treated and untreated subjects: five continuous variables and five dichotomous variables. We assumed that one continuous covariate and one dichotomous covariate had a standardized difference of 0.2 between treated and untreated subjects in the full sample. Similarly, we assumed that the remaining four pairs of continuous and dichotomous variables had standardized differences of 0.3, 0.4, 0.5, and 0.6 between treated and untreated subjects in the full sample. We assumed that the prevalence of exposure was 25%. We randomly generated datasets of size 1000. For each of the 1000 subjects, we randomly generated an exposure status from a Bernoulli distribution with parameter 0.25. For a given standardized difference d, we randomly generated a continuous covariate from the following distribution: $C_i \sim N(T_i \times d, 1)$where T_i denotes the exposure status of the i*th* subject (1 = exposed/treated; 0 = unexposed/untreated). Thus, the distribution of the continuous covariate would be $N(0,1)$ for untreated subjects, and $N(d,1)$ for treated subjects, inducing a standardized difference of d. This was done for $d = 0.2$, 0.3, 0.4, 0.5, and 0.6. These five continuous variables are referred to as C_1, C_2, C_3, C_4, and C_5, respectively. The prevalence of the five dichotomous variables amongst the unexposed subjects was taken to be 0.1, 0.2, 0.3, 0.4, and 0.5. The prevalence of the five dichotomous variables amongst the exposed subjects was selected so that the standardized differences of the five dichotomous variables were 0.2, 0.3, 0.4, 0.5, and 0.6 between treated and untreated subjects. This was achieved by setting the prevalence of the five dichotomous variables amongst the exposed subjects to be 0.168, 0.331, 0.492, 0.642, and 0.776, respectively. These five dichotomous variables are referred to as B_1, B_2, B_3, B_4, and B_5, respectively.

Within each of the 1000 simulated datasets, we estimated the propensity score using a logistic regression model in which an indicator variable denoting treatment status was regressed on the 10 baseline covariates C_1, C_2, C_3, C_4, C_5, B_1, B_2, B_3, B_4, and B_5. We then estimated the weighted conditional standardized differences for each of these 10 covariates using the methods described in 'Weighted conditional standardized differences Section'. The weighted conditional standardized differences were then averaged across the 1000 simulated datasets.

Results

The mean conventional standardized differences comparing the means of the five continuous covariates between treated and untreated subjects in the original samples were 0.20, 0.30, 0.40, 0.50, and 0.60,

DOI: 10.1002/pds

for C_1, C_2, C_3, C_4, and C_5, respectively. The mean conventional standardized differences comparing the prevalence of the five binary covariates between treated and untreated subjects in the original samples were 0.20, 0.30, 0.40, 0.50, and 0.60, for B_1, B_2, B_3, B_4, and B_5, respectively. The mean weighted conditional standardized absolute differences for C_1, C_2, C_3, C_4, and C_5 across the 1000 simulated datasets were 0.067, 0.081, 0.091, 0.115, and 0.138, respectively. The mean weighted conditional standardized absolute differences for B_1, B_2, B_3, B_4, and B_5 across the 1000 simulated datasets were 0.060, 0.067, 0.078, 0.106, and 0.164, respectively. Thus, one observes that conditioning on the estimated propensity score has reduced systematic differences between treated and untreated subjects. The degree of residual differences between treated and untreated subjects conditional on the propensity score increases with the degree of initial differences between treated and untreated subjects in the full (unconditional) sample. It should be noted that none of the weighted conditional standardized absolute differences are zero. Similar results were observed in a prior study looking at the within-quintile balance of observed covariates when stratification on the propensity score was employed.[9] Modest residual imbalance between treated and untreated subjects was observed, particularly in the extreme quintiles. This is in contrast to propensity-score matching in which virtually all imbalance in measured covariates was eliminated.[9]

CASE STUDY

Data sources

We used data on 9107 patients who were discharged alive with an acute myocardial infarction (AMI or heart attack) from 102 hospitals in Ontario, Canada, between 1 April 1999 and 31 March 2001. These data are similar to those reported on elsewhere,[16–18] 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.[19] Data on patient demographics, presenting signs and symptoms, classic cardiac risk factors, co-morbid 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 beta-blocker at hospital discharge, and the outcome of interest was death within 3 years of hospital discharge.

Overall, 6178 (67.8%) of patients received a prescription for a beta-blocker at discharge, while 2929 (32.2%) did not receive a prescription at discharge. Table 1 compares the characteristics of patients who did and did not receive a beta-blocker at hospital discharge. Two sample *t*-tests were used to compare the mean of continuous variables between treated and untreated subjects, while the ^{2} test was used to compare the prevalence of dichotomous risk factors between treated and untreated subjects. Standardized differences were also used to compare the balance in measured baseline covariates between those who did and did not receive a prescription for a beta-blocker at discharge.[3,9,20,21] Twenty-two of the 27 measured baseline covariates had standardized differences that exceeded 0.10.

Propensity score diagnostics

A propensity score model was fit using a logistic regression model in which treatment assignment (beta-blocker *vs.* no beta-blocker) was regressed on the 27 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 beta-blocker at hospital discharge. The estimated propensity scores for treated subjects ranged from 0.0903 to 0.9040, while the estimated propensity scores for untreated subjects ranged from 0.0758 to 0.8927. Non-parametric estimates of the distribution of the propensity score in treated and untreated subjects are illustrated in Figure 1. While the distribution of the propensity score in untreated subjects had a heavier left tail, the support of the distribution of the estimated propensity score was similar in treated and untreated subjects. Since for each untreated subject, there was a treated subject with a similar propensity score, no treated or untreated subjects were excluded from the analysis. This reflects how covariate adjustment using the propensity score is typically employed in practice: all subjects are retained for the analysis. Occasionally, those treated subjects with propensity scores greater than those of all untreated subjects are excluded. Similarly, untreated subjects with propensity scores lower than those of all treated subjects are also excluded. Given the similarity of the support of the distribution of the estimated propensity score between treated and untreated subjects, this was not done in our case study.

The methods described in Section 'Balance diagnostics for the propensity score model', were

DOI: 10.1002/pds

Table 1. Baseline characteristics of treated and untreated subjects

Variable	Beta-blocker: No ($N = 2929$)	Beta-blocker: Yes ($N = 6178$)	p-value	Unconditional standardized difference	Weighted conditional standardized difference
Demographic characteristics					
Age	69.6 ± 13.5	65.0 ± 13.3	<.001	0.342	0.221
Female	1144 (39.1%)	1984 (32.1%)	<.001	0.147	0.087
Presenting signs and symptoms					
Acute CHF/Pulmonary edema	214 (7.3%)	224 (3.6%)	<.001	0.173	0.072
Classic cardiac risk factors					
Diabetes	842 (28.7%)	1494 (24.2%)	<.001	0.105	0.046
Current smoker	916 (31.3%)	2158 (34.9%)	<.001	0.077	0.041
CVA/TIA	354 (12.1%)	493 (8.0%)	<.001	0.142	0.115
Hyperlipidemia	767 (26.2%)	2132 (34.5%)	<.001	0.179	0.118
Hypertension	1343 (45.9%)	2793 (45.2%)	0.565	0.013	0.019
Family history of CAD	745 (25.4%)	2195 (35.5%)	<.001	0.217	0.105
Co-morbid conditions					
Angina	975 (33.3%)	1982 (32.1%)	0.251	0.026	0.044
Cancer	110 (3.8%)	154 (2.5%)	<.001	0.075	0.035
Dementia	142 (4.8%)	134 (2.2%)	<.001	0.157	0.063
Previous AMI	739 (25.2%)	1314 (21.3%)	<.001	0.095	0.045
Asthma	323 (11.0%)	181 (2.9%)	<.001	0.359	0.019
Depression	256 (8.7%)	377 (6.1%)	<.001	0.104	0.056
Peripheral vascular disease	281 (9.6%)	369 (6.0%)	<.001	0.141	0.068
Chronic CHF	189 (6.5%)	177 (2.9%)	<.001	0.183	0.099
Vital signs on admission					
Systolic BP	146.8 ± 31.4	149.9 ± 30.9	<.001	0.102	0.080
Diastolic BP	81.8 ± 18.6	84.9 ± 18.3	<.001	0.173	0.157
Heart rate	86.9 ± 25.9	82.1 ± 22.7	<.001	0.204	0.008
Respiratory rate	22.2 ± 6.5	20.3 ± 4.8	<.001	0.351	0.022
Laboratory tests					
Glucose	9.8 ± 5.2	9.2 ± 5.2	<.001	0.118	0.001
White blood count	10.6 ± 5.5	10.0 ± 4.3	<.001	0.130	0.010
Hemoglobin	135.2 ± 20.0	140.2 ± 17.7	<.001	0.267	0.134
Sodium	138.7 ± 4.2	139.2 ± 3.5	<.001	0.111	0.043
Potassium	4.1 ± 0.6	4.1 ± 0.5	<.001	0.127	0.038
Creatinine	114.2 ± 77.4	98.8 ± 50.3	<.001	0.254	0.008

Note: Continuous variables are represented as Mean ± Standard deviation, while dichotomous variables are represented as N (%).

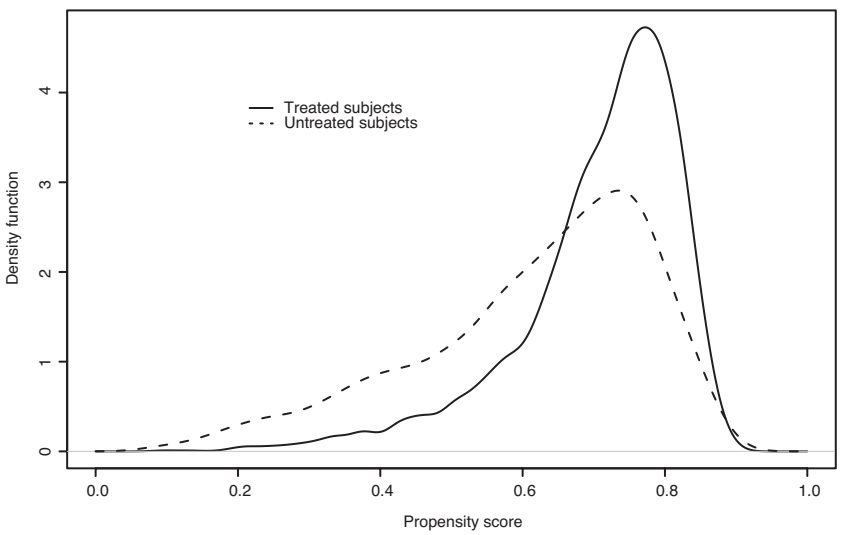

Figure 1. Distribution of the propensity score in treated and untreated subjects

DOI: 10.1002/pds

then used to assess whether, conditional on the estimated propensity score, treated and untreated subjects had similar distributions of both continuous and dichotomous baseline covariates.

The weighted conditional standardized differences for each of the measured baseline covariates are reported in the rightmost column in Table 1. The weighted conditional standardized differences exceeded 0.10 for 5 of the 27 baseline covariates. The weighted conditional standardized differences for age, history of CVA/TIA, history of hyperlipidemia, diastolic blood pressure, and haemoglobin were 0.221, 0.115, 0.118, 0.157, and 0.134, respectively. In contrast, the crude standardized differences exceeded 0.10 for 22 of the 27 baseline covariates. While there is no threshold for standardized differences that has been uniformly accepted as indicative of meaningful imbalance, authors of several papers have suggested that a standardized difference of 0.1 may indicate potentially meaningful imbalance in a covariate between treated and untreated subjects.[3,8,9,20,21]

The prevalence of each of the 16 dichotomous variables, conditional on the propensity score, in treated and untreated subjects separately is displayed in Figures 2 and 3. The relationship between the propensity score and the prevalence of the risk factor is depicted by a solid line for patients prescribed a beta-blocker and by a dashed line for patients not prescribed a beta-blocker. Along the base of each plot, a jittered rug plot of the estimated propensity scores has been included to describe the distribution of the propensity score in the sample. For some variables, such as family history of coronary artery disease and asthma, the prevalence of the risk factor, conditional on the propensity score, tended to be similar between treated and untreated subjects. However, for the majority of risk factors (e.g., female gender, acute CHF/pulmonary edema at admission, history of CVA/TIA, history of hypertension, history of angina, history of cancer, dementia, previous AMI, depression, peripheral vascular disease, and chronic congestive heart failure), the prevalence of these variables, conditional on the propensity score, was different between treated and untreated subjects. In particular, the difference in prevalence between treated and untreated subjects was amplified at low values of the propensity score, and was often minimal at higher values of the propensity score. Interestingly, one of the variables (hypertension) was balanced between treated and untreated subjects in the overall sample (standardized difference of 0.013 and a p-value of 0.565 for the ^{2} test), but, conditional on the propensity score, was imbalanced between treated and untreated subjects. However, the weighted conditional standardized difference was 0.019, indicating that overall, conditional on the propensity score, there was a similar prevalence of hypertension between treated and untreated subjects.

The distribution of the 11 continuous covariates in treated and untreated subjects, conditional on the estimated propensity score, is described in Figures 4 and 5. The conditional distribution of a specific covariate in a particular treatment group is described by five lines, representing the 5th, 25th, 50th, 75th, and 95th percentiles of the conditional distribution. The 50th percentile is represented by a thick line, the 25th and 75th percentiles by medium lines, and the 5th and 95th percentiles by thin lines. The conditional distribution of the covariates for treated patients is depicted in solid red lines and for untreated patients by dashed blue lines. Some variables (e.g., white blood count, glucose, and potassium) had conditional distributions that were very similar between treated and untreated subjects. However, for other variables, the conditional distributions differed between treated and untreated subjects. Among subjects with a low propensity score, the median estimated age for treated subjects was greater than the 95th percentile of age for untreated subjects. Similarly, among subjects with a low propensity score, the percentiles of the conditional distribution of diastolic blood pressure in untreated subjects were higher than the corresponding percentiles of the conditional distribution in treated subjects. However, the conditional distributions were comparable among subjects with high propensity scores. For creatinine, the conditional distribution exhibited greater positive skewness in untreated subjects than it did in treated subjects.

DISCUSSION

In the current paper, we have proposed methods to assess the adequacy of the specification of the propensity score model. Covariate adjustment using the propensity score assumes that, conditional on the propensity score, the distribution of measured baseline covariates is similar between treated and untreated subjects. Importantly, it is assumed that this is true over the entire distribution of the estimated propensity score. We developed the weighted conditional standardized absolute difference to quantitatively compare the conditional difference in baseline covariates between treated and untreated subjects. We also proposed that quantile regression models be used to qualitatively examine the conditional distribution

DOI: 10.1002/pds

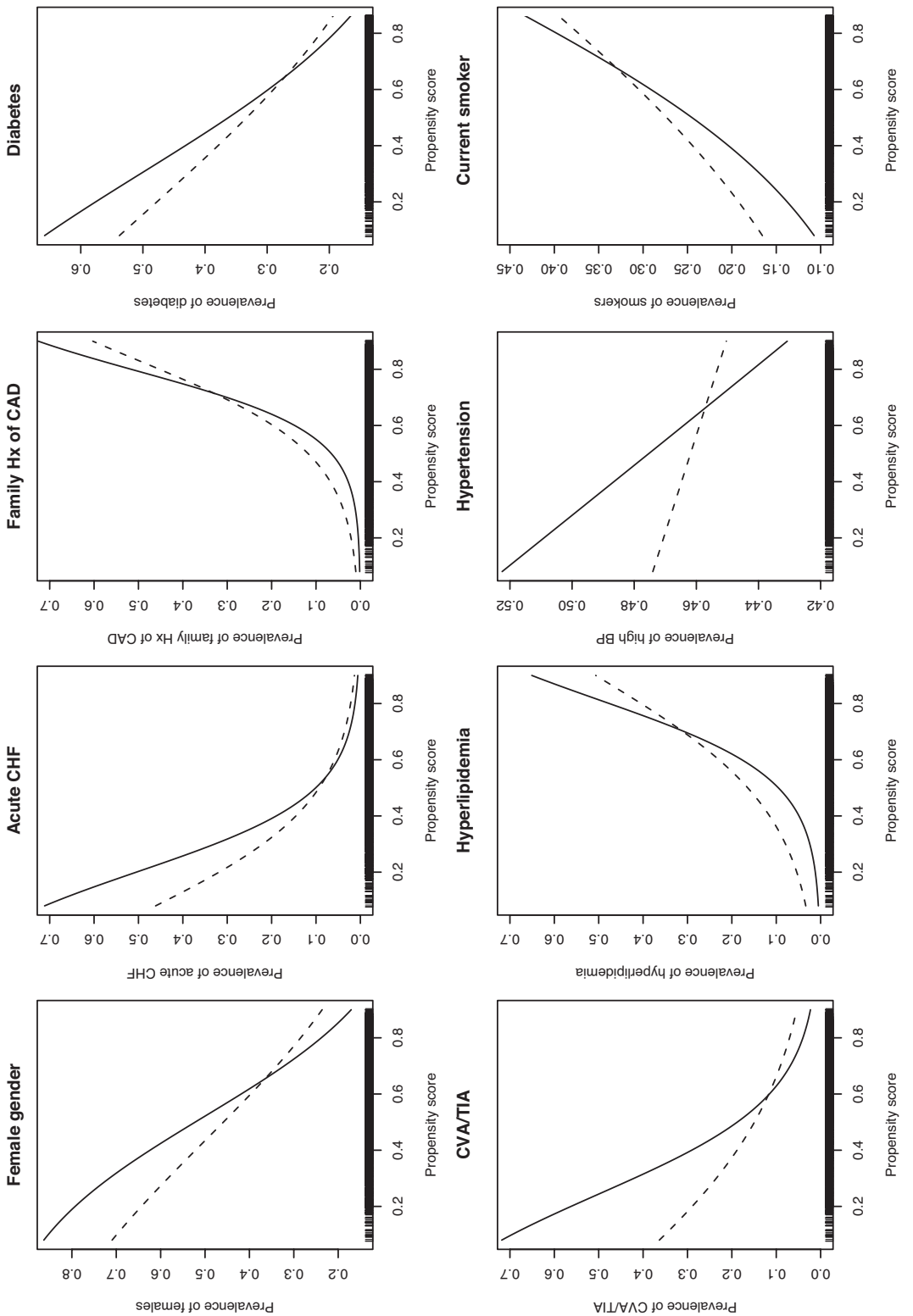

Figure 2. Conditional prevalence of risk factors

DOI: 10.1002/pds

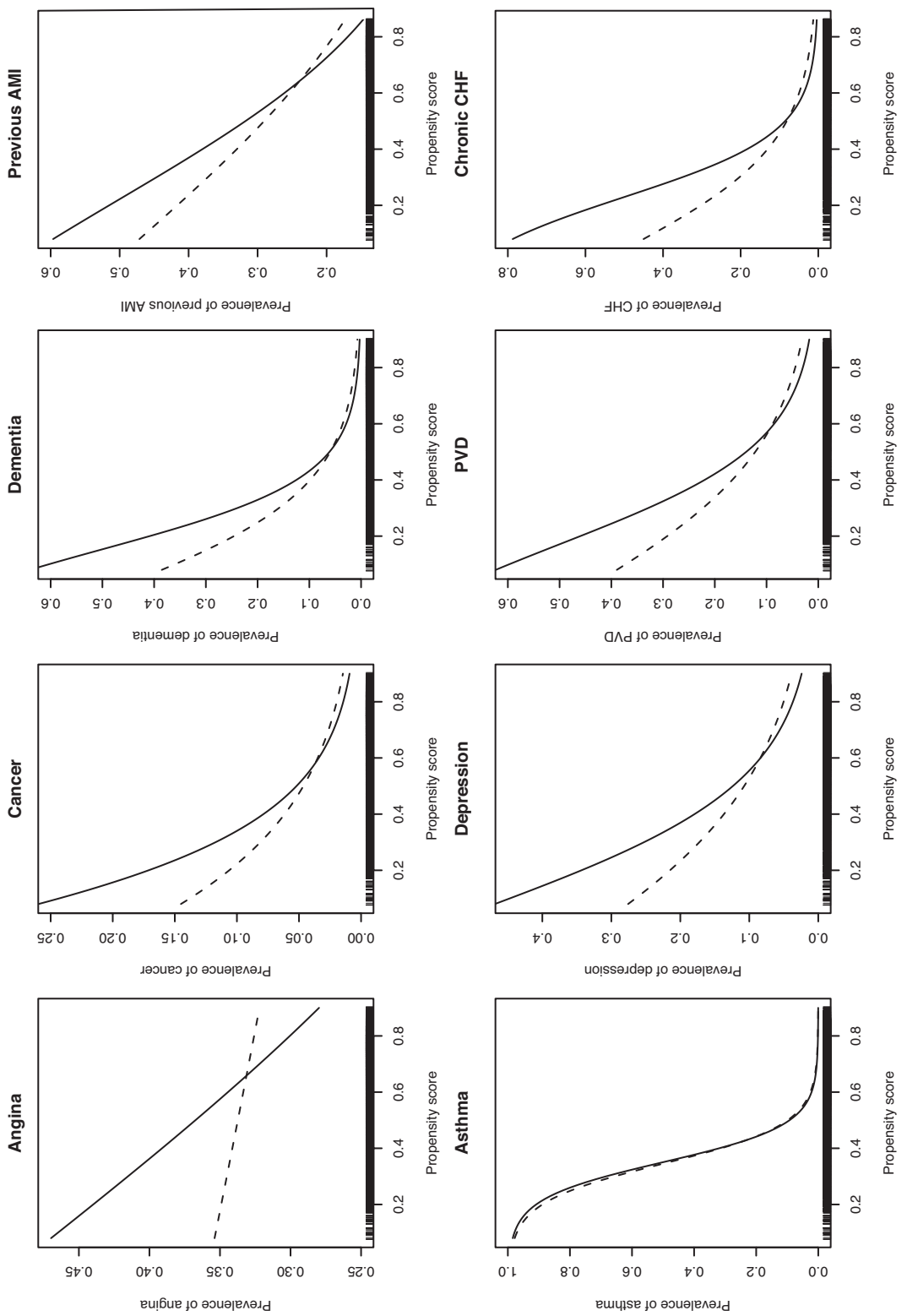

Figure 3. Conditional prevalence of risk factors

DOI: 10.1002/pds

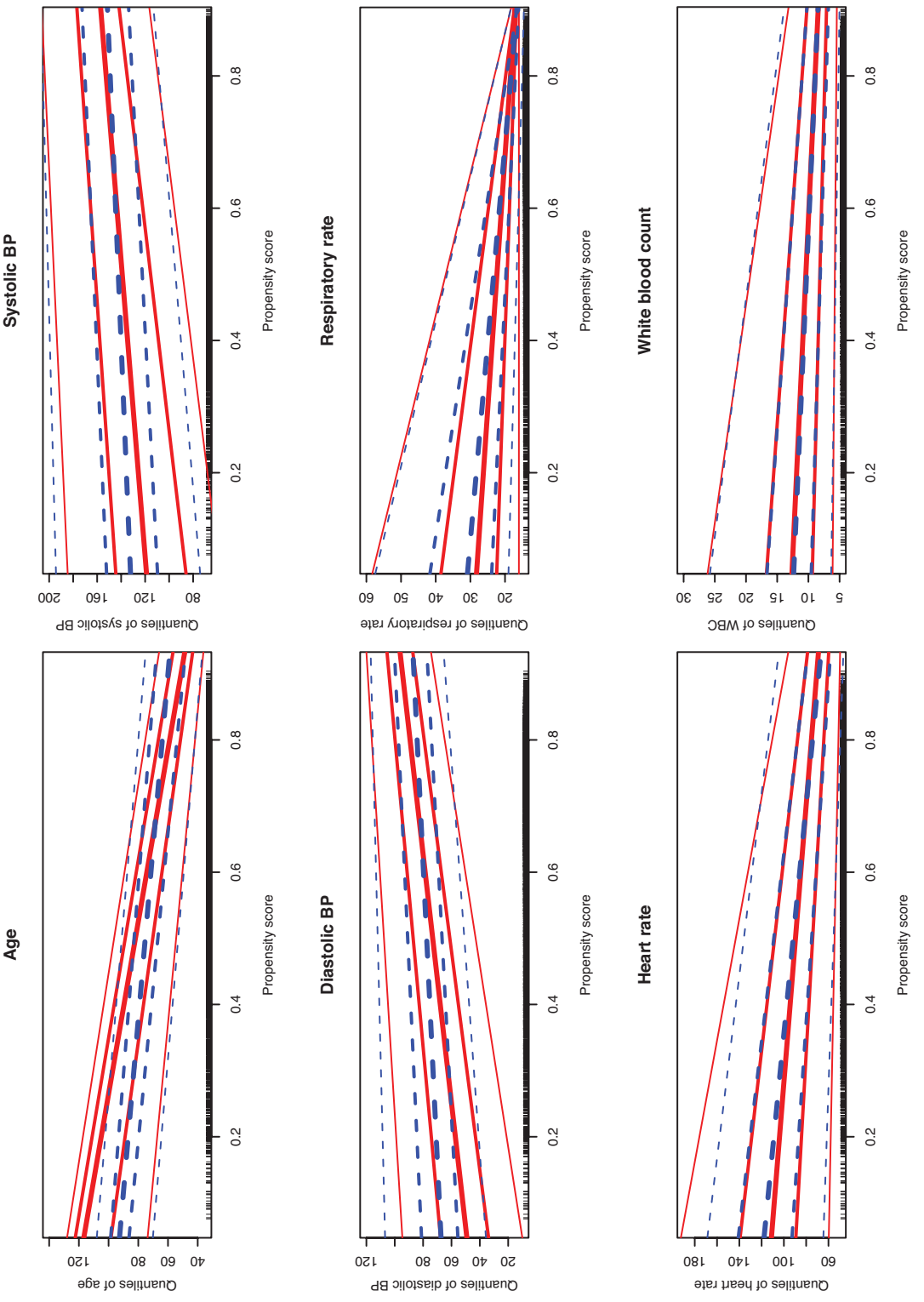

Figure 4. Conditional distribution of covariates

DOI: 10.1002/pds

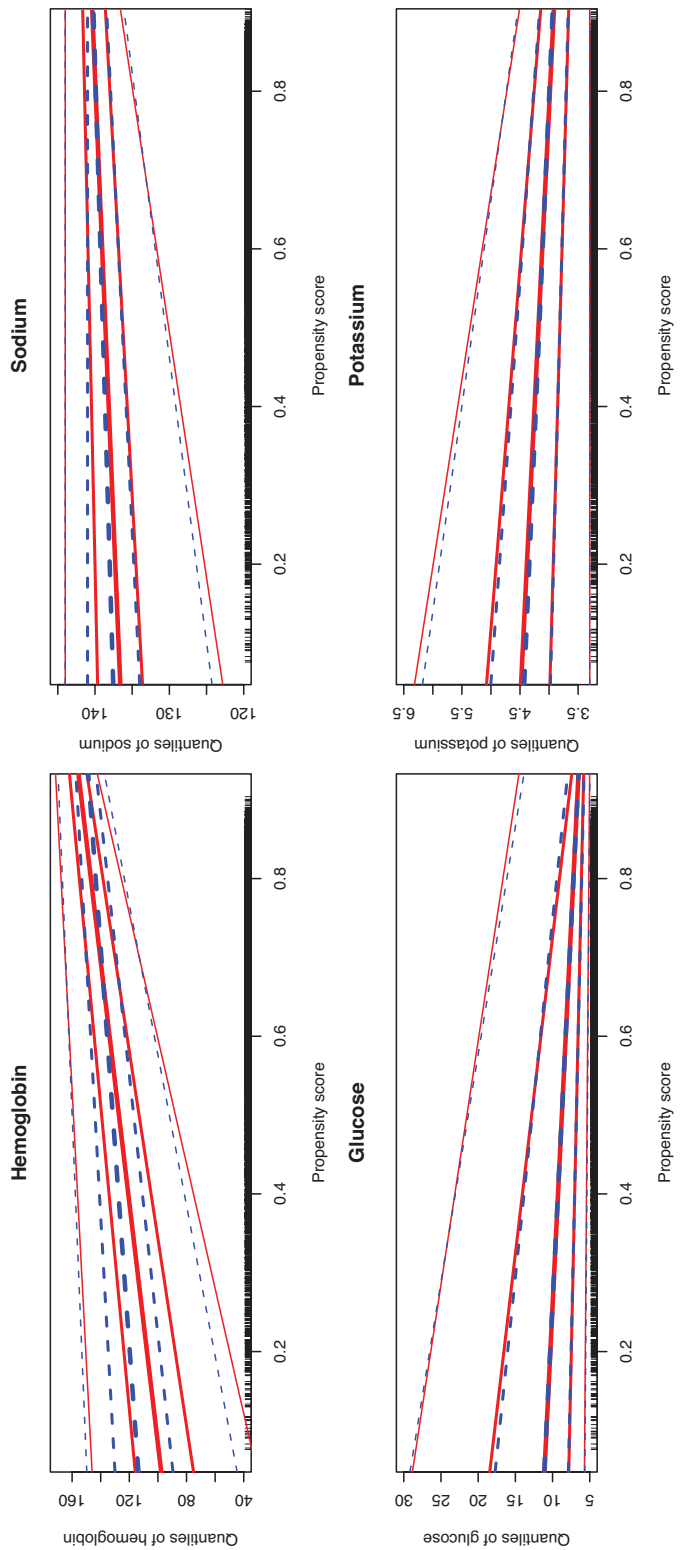

Figure 5. Conditional distribution of covariates

DOI: 10.1002/pds

of continuous baseline covariates between treated and untreated subjects.

Methods for assessing balance have been developed for use when stratifying on the quintiles of the propensity score and when matching on the propensity score. However, no comparable balance diagnostics have been proposed for use when using covariate adjustment using the propensity score. Covariate adjustment using the propensity score was proposed by Rosenbaum and Rubin[1] in their original article on the propensity score. Furthermore, it is the most commonly employed propensity-score method in the medical literature.[4–6] The diagnostics described in this paper will allow for a more rigorous implementation of covariate adjustment using the propensity score. In the current paper, we have not addressed estimation using covariate adjustment using the propensity score, as this is addressed elsewhere.[1,22–24] The focus of the current manuscript is balance diagnostics for when covariate adjustment using the propensity score is employed.

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 [25]. 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 beta-blocker at discharge) and measured baseline covariates. Rubin suggests that '*diagnostics for the successful design of observational studies based on estimated propensity scores ... is a critically important activity in most observational studies*' [26]. Our proposed diagnostics will thus contribute to improving study design when the propensity score is used for covariate adjustment. While covariate adjustment using the propensity score was described by Rosenbaum and Rubin[1] and is the most commonly used propensity-score method in the medical literature, it is not without its limitations. In particular, unlike other propensity-score methods, it assumes that the outcomes regression model has been correctly specified[26].

In summary, we have proposed diagnostics for assessing whether the propensity score model has been adequately specified when covariate adjustment using the propensity score is used to estimate causal treatment effects. Methods have previously been developed for assessing whether matching or stratifying on the propensity score has reduced or eliminated systematic differences between treated and untreated subjects. However, no comparable methods have been proposed for when the propensity score is used for covariate adjustment. Given that this approach is the most commonly implemented propensity-score method in the medical literature, the proposed methods will improve the implementation of propensity score methods in the medical literature.

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 subjects with similar values of the propensity score.
- In the context of covariate adjustment using the propensity score, the weighted conditional standardized difference can be used to determine whether, conditional on the propensity score, the mean of observed baseline covariates is similar between treated and untreated subjects.
- Quantile regression can be used to compare whether, conditional on the propensity score, the distribution of continuous baseline covariates is similar between treated and untreated subjects.

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. 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. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). The EFFECT data used in the study 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. The study was approved by the Research Ethics Board of Sunnybrook Health Sciences Centre.

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, 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.
4. 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.
5. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods give similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58**: 550–559.
6. 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**: 437–447.
7. 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.
8. 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.
9. 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.
10. 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.
11. Flury BK, Riedwyl H. Standard distance in univariate and multivariate analysis. *Am Stat* 1986; **40**: 249–251.
12. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc Ser A (Stat Soc)* 2008; **171**: 451–502.
13. Austin PC, Tu JV, Daly PA, Alter DA. The use of quantile regression in health care research: a case study examining gender differences in the delivery of thrombolysis. *Stat Med* 2005; **24**: 791–816.
14. Austin PC, Schull MJ. Quantile regression: a statistical tool for out-of-hospital research. *Acad Emerg Med* 2003; **10**: 789–797.
15. Austin PC. The performance of different propensity-score matching methods used in the medical literature. *Biom J Manuscript in-press.*
16. 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.
17. 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.
18. 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.
19. Tu JV, Donovan LR, Lee DS, *et al*. *Quality of Cardiac Care in Ontario*. Institute for Clinical Evaluative Sciences: Toronto, Ontario, 2004.
20. Austin PC. A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Stat Med* 2008; **27**: 2037–2049.
21. 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.
22. Austin PC, Grootendorst P, Normand SLT, 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**: 754–768.
23. Austin PC. The performance of different propensity score methods for estimating relative risks. *J Clin Epidemiol* 2008; **61**: 537–545.
24. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med* 2007; **26**: 3078–3094.
25. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol* 2001; **2**: 169–188.
26. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiol Drug Saf* 2004; **13**: 855–857.

APPENDIX. EXAMPLE OF A SETTING IN WHICH TREATED AND UNTREATED SUBJECTS HAVE DIFFERENT DISTRIBUTION OF A BASELINE VARIABLE BUT IN WHICH THERE IS ACCEPTABLE WITHIN-STRATUM BALANCE

Assume that the first quintile of the estimated propensity scores includes those subjects whose estimated propensity score lies in the interval (0,0.2).

Let X_T and X_C denote the value of the baseline variable X for treated and untreated subjects, respectively. Let Z denote the value of the estimated propensity score.

Define $X_T = 0.1Z + \varepsilon$ and $X_C = 0.2 - 1.9Z + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$. This is equivalent to saying that within the first stratum, $X_T \sim N(0.1Z, \sigma)$ and that $X_C \sim N(0.2 - 1.9Z, \sigma)$. Then, for $Z = 0.1$ (the midpoint of the stratum), the distribution of X will be the same for treated and untreated subjects. However, for all other values of Z in the interval (0,0.2), the distribution of X will be different for treated and untreated subjects.

Assume within the propensity-score stratum (0,0.2), the propensity score is uniformly distributed for both treated and untreated subjects. Then, one can determine the mean value of X_T and X_C within the propensity-score stratum (note: E[X] denotes the mean or expectation of a random variable):

$$E[X_T] = E[0.1Z + \varepsilon] = 0.1E[Z] + E[\varepsilon] = 0.1 \times 0.1 + 0 = 0.01.$$

Similarly,

DOI: 10.1002/pds

$$E[X_Y] = E[0.2 - 1.9Z + \varepsilon] = E[0.2] - 1.9E[Z] + E[\varepsilon] = 0.2 - 1.9 \times 0.1 + 0 = 0.2 - 0.19 = 0.01$$

Therefore, the average value of X is the same between treated and untreated subjects within the first propensity-score stratum. However, for all values of the propensity score in the stratum (except for $Z = 0.1$), the distribution of X is different between treated and untreated subjects.

The above example is intended to examine to highlight a specific setting. However, each of the assumptions can be relaxed, and the same principal can still be found to hold. For simplicity of the mathematical derivations we have presented a simple scenario.

DOI: 10.1002/pds