STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2084–2106
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2328

A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use

Peter C. Austin[1,2,3,*,†] and Muhammad M. Mamdani[1,3,4]

[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, Canada*
[4]*Faculty of Pharmacy, University of Toronto, Canada*

SUMMARY

There is an increasing interest in the use of propensity score methods to estimate causal effects in observational studies. However, recent systematic reviews have demonstrated that propensity score methods are inconsistently used and frequently poorly applied in the medical literature. In this study, we compared the following propensity score methods for estimating the reduction in all-cause mortality due to statin therapy for patients hospitalized with acute myocardial infarction: propensity-score matching, stratification using the propensity score, covariate adjustment using the propensity score, and weighting using the propensity score. We used propensity score methods to estimate both adjusted treated effects and the absolute and relative risk reduction in all-cause mortality. We also examined the use of statistical hypothesis testing, standardized differences, box plots, non-parametric density estimates, and quantile–quantile plots to assess residual confounding that remained after stratification or matching on the propensity score. Estimates of the absolute reduction in 3-year mortality ranged from 2.1 to 4.5 per cent, while estimates of the relative risk reduction ranged from 13.3 to 17.0 per cent. Adjusted estimates of the reduction in the odds of 3-year death varied from 15 to 24 per cent across the different propensity score methods. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: propensity score; pharmacoepidemiology; acute myocardial infarction; statins; statistical methods

1. INTRODUCTION

There is an increasing interest in using observational data to assess the impact of medical treatment or therapy on health outcomes. While randomized controlled trials (RCTs) are

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

Contract/grant sponsor: Ontario Ministry of Health and Long Term Care
Contract/grant sponsor: Canadian Institutes of Health Research (CIHR)

Received 29 November 2004
Accepted 26 May 2005

considered the gold standard for assessing the effectiveness of therapy, there are instances in which these are not feasible for either ethical or practical reasons. Furthermore, patients enrolled in clinical trials may not be representative of the population to which the therapy will eventually be applied. The use of observational data allows investigators to estimate the effectiveness of treatment in the general population.

In RCTs, randomization ensures that the different arms of the trial tend to be balanced, both in terms of measured and unmeasured characteristics of the population. Thus, the only systematic difference between treated and untreated patients is the exposure to treatment. This allows one to obtain unbiased estimates of the average treatment effect within the context of the study design. In observational studies the distribution of subject characteristics is likely to vary systematically between different treatment groups since treatment is not randomly assigned but rather passively observed and is often related to patient prognosis, patient characteristics, and physician preference. Thus, confounding may occur.

There is a growing interest in the use of propensity score-based methods for estimating treatment effects in observational studies. The propensity score is defined as a subject's probability of treatment assignment conditional on measured covariates [1–3]. Three propensity score-based methods are commonly used in the medical literature: matching, covariate adjustment, and stratification or subclassification [3]. Readers are referred elsewhere for overviews and more theoretical discussion of the propensity score [1–10]. There is no consensus in the clinical literature as to which method is preferable. Furthermore, there is only a limited awareness of the relative strengths and limitations of each propensity score method. Two recent systematic reviews on the use of propensity score methods in the clinical literature revealed that the most commonly used method in the clinical literature was covariate adjustment using the propensity score [11, 12]. Furthermore, it was documented that propensity score methods are often poorly implemented in applied clinical research.

The purpose of the current study was two-fold. First, to compare estimates of treatment effectiveness obtained using different propensity score methods. In doing so, the relative merits and limitations of each method will be highlighted. Second, to carry out a detailed propensity score analysis, which will serve as a model for clinical investigators who wish to implement propensity score methods. As a test case, we examine the effect of statin lipid-lowering therapy on reducing all-cause mortality for patients discharged alive from hospital with a diagnosis of acute myocardial infarction (AMI).

2. METHODS

2.1. Data sources

Detailed clinical data were collected on a sample of 11 524 patients discharged from Ontario hospitals between April 1, 1999 and March 31, 2001 by retrospective chart review. These data were collected as part of the Enhanced Feedback For Effective Cardiac Treatment (EFFECT) study, an ongoing initiative intended to improve the quality of care for patients with cardiovascular disease in Ontario [13]. Data on patient history, cardiac risk factors, comorbid conditions and vascular history, vital signs, and laboratory tests were collected for this sample. Furthermore, data on medications prescribed at discharge were available for each patient. Linking patients to the registered persons database (RPDB) using encrypted health card

numbers allowed us to determine each patient's vital status. We allowed each patient to have 3 years of follow-up post-discharge. Patients who were missing data on important vital signs at admission or laboratory values were excluded from all subsequent analyses.

Differences in measured characteristics between treated and untreated patients were assessed using two methods. First, the statistical significance of the difference in either the proportion of patients having a dichotomous risk factor or the difference in the mean of a continuous covariate between treated and untreated patients was tested. A chi-square test was used for dichotomous variables and a *t*-test was used for continuous variables. Second, the standardized difference was computed for each covariate. The standardized difference is defined as

$$d = \frac{100(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{(s^2_{\text{treatment}} + s^2_{\text{control}})/2}} \quad \text{for continuous variables} \tag{1}$$

and by

$$d = \frac{100(p_{\text{treatment}} - p_{\text{control}})}{\sqrt{[p_T(1-p_T) + p_C(1-p_C)]/2}} \quad \text{for dichotomous variables} \tag{2}$$

It has been suggested that a standardized difference of greater than 10 per cent represents meaningful imbalance in a given covariate between treatment groups [14].

In the current study, we chose to focus on estimating the treatment effects for statins on reducing all-cause mortality post-AMI since there is evidence from meta-analyses of randomized clinical trials that this therapy is associated with reduced post-AMI mortality [15, 16].

2.2. Propensity score development

Clinical data contained the following variables that were potential confounders of the treatment effect: demographic characteristics: age, gender; presenting signs and symptoms: shock and acute congestive heart failure (acute CHF)/pulmonary oedema; classical cardiac risk factors: diabetes, CVA/TIA (history of cerebrovascular accident or transient ischaemic attack), history of hyperlipidaemia, hypertension, family history of heart disease, and smoking history; comorbid conditions: angina, cancer, congestive heart failure (CHF)/pulmonary oedema, and renal disease; vital signs on admission: systolic and diastolic blood pressure, heart rate, and respiratory rate; laboratory test results (haematology): haemoglobin (Hgb), white blood count (WBC); and laboratory test results (chemistry): sodium, potassium, glucose, and creatinine.

A clinical model had previously been developed for predicting 30-day AMI mortality using detailed clinical data [17]. This model adjusts for age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission. This model had an area under the ROC curve of 0.822 for predicting 30-day AMI mortality. This model will serve as the initial step in constructing the propensity score model.

We developed a propensity score model to predict the probability that the patient would be given a prescription for a statin at hospital discharge. We used a structured, iterative approach, similar to that described by Rosenbaum and Rubin [1], to balance the measured covariates. The propensity score model was developed to balance the distribution of the above possible confounders between treated and untreated patients within each quintile of the estimated propensity score. The iterative algorithm for developing the propensity score allowed for the

inclusion of main effects, interaction terms, in addition to quadratic and cubic terms for continuous variables. Following the derivation of the propensity score model, untreated patients who had a lower estimated propensity score than any treated patient were excluded from subsequent analyses. The coefficients of the propensity model were then re-estimated on this reduced data set. This decision was made because of the belief that the excluded patients were different from all treated patients, and may thus have not been candidates for statin therapy.

2.3. Estimation of treatment effect—adjusted estimates

Once the propensity score model had been developed, we sought to estimate the reduction in all-cause mortality attributable to the post-discharge use of statins. This was done in four different fashions: stratifying on the quintiles of the estimated propensity score, matching treated and untreated patients using the estimated propensity score, covariate adjustment using the estimated propensity score, and weighting using the propensity score. We describe each of these methods below. In what follows, let p_{ij} denote the probability of 3-year mortality for the ith patient in the jth quintile and p_i denote the probability of 3-year mortality for the ith patient (either in the entire cohort or in the matched sample); let Z_{ij} or Z_i denote treatment assignment for this patient; let e_{ij} or e_i denote the estimated propensity score for this patient; let X_{ij} or X_i denote a vector of observed covariates for this patient; and let β denote a vector of regression coefficients for the covariates in the vector X_{ij} or X_i.

2.3.1. Stratifying on the quintiles of the propensity score. Patients were stratified into quintiles according to their predicted probability of receiving a statin prescription at hospital discharge. Cochran demonstrated that stratifying on the quintiles of a continuous confounding variable could be expected to eliminate approximately 90 per cent of the bias to an imbalance in the confounding variable between two groups [18]. Rosenbaum and Rubin extended this result to include stratifying on the quintiles of the propensity score could be expected to eliminate approximately 90 per cent of the bias due to an imbalance in measured covariates between the two groups [1]. The degree to which the estimated propensity score balanced the measured covariates was assessed by comparing the distribution of the propensity score between treated and untreated subjects within each quintile of the propensity score. Furthermore, within each quintile, the standardized difference of each covariate was computed.

Within each quintile, a univariate logistic regression model was used, in which the probability of mortality within 3 years of discharge was modelled using exposure to statins as the sole predictor variable. From each quintile-specific regression model, an estimated log-odds ratio was obtained. A pooled average treatment effect was obtained by combining the estimates across quintiles. This method was repeated using a stratum-specific logistic regression model that estimated the effect of statin on the reduction in mortality after adjusting for the eight predictor variables in our initial risk adjustment model (age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission). These methods can be described mathematically as follows:

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j} Z_{ij}, \quad j = 1, \ldots, 5$$

$$\beta = \frac{1}{5} \sum_{j=1}^{5} \beta_{1j}$$

Then $\exp(\beta)$ is the pooled estimated treated effect. This method can then be repeated with the stratum specific regression model replaced by

$$\text{logit}(p_{ij}) = \beta_{0j} + \beta_{1j}Z_{ij} + \beta_{2j}X_{ij}, \quad j = 1, \ldots, 5$$

2.3.2. Matching on the propensity score. The second propensity score method that we used was a matched-paired analysis. For each subject, we computed the logit of the estimated propensity score. We then used a greedy-matching algorithm to match subjects using calipers that were defined to have a maximum width of 0.2 standard deviations of the logit of the estimated propensity score [19]. The degree to which matching on the propensity score created a matched sample that balanced measured covariates between treated and untreated patients was assessed in two methods. First, using either paired *t*-tests or McNemar's test, the mean or prevalence of each covariate was compared between treated and untreated patients. Second, standardized differences between treated and untreated patients were computed for each covariate.

The treatment effect was estimated by fitting a logistic regression model to estimate the impact of statin therapy on the reduction in the odds of mortality, stratifying on the matched pairs. This method of estimating the treatment effect can be described as

$$\text{logit}(p_i) = \beta_0 + \beta_1 Z_i$$

The logistic regression model can be estimated using generalized estimating equation (GEE) methods to incorporate the matched-pairs design. Then $\exp(\beta_1)$ is the estimated treated effect using the matched sample.

2.3.3. Covariate adjustment using the propensity score. The third method used the estimated propensity score as a covariate in a multivariate logistic regression model. Mortality was regressed on the estimated propensity score and a dichotomous variable denoting receipt of a prescription for a statin at discharge. Direct modelling of the outcome assumes that the regression model has been correctly specified. Generally, applied investigators have only included a linear term for the estimated propensity score. We examined the appropriateness of this practice in our example in two different methods. First, we included both linear and quadratic terms for the propensity score in the regression model. Second, we accounted for possibly non-linearity in the relationship between mortality and the propensity score by using restricted cubic splines with five knots to adjust for the propensity score [20]. Finally, we repeated this method with age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission added to the logistic regression model in addition to the propensity score and the exposure variable. This method can be described as

$$\text{logit}(p_i) = \beta_0 + \beta_1 Z_i + \beta_2 e_i$$

Then $\exp(\beta_1)$ is the estimated treated effect using the propensity score for covariate adjustment. The above logistic regression model can be modified to either incorporate quadratic terms for the propensity score, or to model the propensity score as a cubic spline.

2.3.4. Propensity score weighting. A recently proposed class of estimators, known as weighted estimators [21], are part of a family of estimators proposed by Robins *et al.* [22]. They use inverse probability weighting in conjunction with regression modelling to obtain estimates of treatment effect. An application of regression modelling with weights derived from the propensity score is described by Joffe *et al.* [23]. Rosenbaum was one of the first to propose methods that used propensity score-based weighting [24]. These methods have been rarely used in the clinical literature, despite having attractive theoretical properties. Let $\hat{e}_i$ and Z_i, denote the estimated propensity score and the treatment assignment indicator, respectively, for the ith subject. Then the weights are defined to be

$$w_i = \frac{Z_i}{\hat{e}_i} + \frac{1 - Z_i}{1 - \hat{e}_i}$$

We used propensity score weighting with two different regression models. First, we used logistic regression to regress 3-year mortality on statin exposure at discharge and the following predictors of AMI mortality: age, cardiogenic shock, systolic blood pressure, glucose, family history of heart disease, white blood count, creatinine levels, and respiratory rate. Confidence intervals were obtained using robust variance estimates, as suggested elsewhere [23].

In order to build a more complex regression model, we used classical model building techniques to construct a predictive model for 3-year mortality. This was done using the following steps sequentially. First, we estimated a series of univariate logistic regression models, in which the univariate association between each variable listed in Table I and 3-year mortality was determined. Those variables that were significant at a significance level of 0.05 were retained for possible inclusion in the final model. Second, for each continuous variable, the fit of a univariate model with only a simple linear term was compared with that of a model that incorporated a restricted cubic spline function for that continuous variable. Variables that represented evidence of a linear relationship with mortality were represented linearly in subsequent steps, while cubic splines were used to represent the remaining continuous variables in subsequent steps. Third, we examined potential interactions between predictor variables. Due to the large number of potential interactions, we consulted with two clinical subject-matter experts to derive a list of *a priori* clinically plausible interactions. The statistical significance of each of these *a priori* specified interactions was then tested in our data, and those that were found to be significant were retained for subsequent steps. Fourth, we then fit a complete model including all main effects, interactions, and cubic-spline representations for continuous variables that were identified in the second step. Fifth, using a series of likelihood ratio tests, we then compared whether a cubic spline was still necessary in the full model, to represent those variables that had been indicated in the second step to have a non-linear relationship with mortality. Those variables for which there was now evidence of a linear relationship were now entered as such in the full model. We then used a series of likelihood ratio tests to test the significance of interactions that involved covariates represented as cubic splines. Non-significant interactions were removed from the model. Finally, non-significant interaction involving simple terms were sequentially eliminated until all non-significant interactions were eliminated from the model. This final model, which will be described as a complex predictive model, was then used to estimate the treatment effect of statins. This method can be described as

$$\text{logit}(p_i) = \beta_0 + \beta_1 Z_i + \beta X_{ij}$$

Table I. Comparisons of statin users and non-users.

Characteristic	Statin non-users $N = 6055$	Statin users $N = 3049$	P-Value
Demographic characteristics			
Age	68.11 ± 13.85	63.36 ± 12.39	<0.001
Female	2241 (37.0%)	887 (29.1%)	<0.001
Presenting characteristics			
Shock	46 (0.8%)	12 (0.4%)	0.038
Acute CHF/pulmonary oedema	316 (5.2%)	122 (4.0%)	0.010
AMI risk factors			
Family history of CAD	1762 (29.1%)	1177 (38.6%)	<0.001
Diabetes	1561 (25.8%)	774 (25.4%)	0.684
CVA/TIA	610 (10.1%)	237 (7.8%)	<0.001
Hyperlipidaemia	1138 (18.8%)	1761 (57.8%)	<0.001
High BP	2681 (44.3%)	1453 (47.7%)	0.002
Current smoker	2004 (33.1%)	1070 (35.1%)	0.057
Comorbidities			
Angina	1869 (30.9%)	1086 (35.6%)	<0.001
Cancer	191 (3.2%)	73 (2.4%)	0.041
CHF	275 (4.5%)	91 (3.0%)	<0.001
Renal disease	34 (0.6%)	13 (0.4%)	0.396
Vital signs on admission			
Systolic BP	148.69 ± 31.57	149.35 ± 30.07	0.338
Diastolic BP	83.62 ± 18.62	84.49 ± 18.00	0.033
Heart rate	84.60 ± 24.31	81.71 ± 22.96	<0.001
Respiratory rate	21.20 ± 5.74	20.30 ± 4.78	<0.001
Laboratory values			
White blood count	10.34 ± 4.87	10.03 ± 4.42	0.003
Haemoglobin	137.54 ± 19.35	140.64 ± 16.92	<0.001
Sodium	138.92 ± 3.92	139.22 ± 3.29	<0.001
Glucose	9.43 ± 5.11	9.24 ± 5.30	0.092
Potassium	4.10 ± 0.57	4.07 ± 0.51	0.006
Creatinine	105.71 ± 65.45	99.89 ± 50.03	<0.001

Note: Continuous variables are reported as mean±standard deviation, while dichotomous variables are reported as number with condition (per cent).

The logistic regression model is estimated using conventional maximum likelihood methods, and each subject was weighted by the weights w_i described above. Then $\exp(\beta_1)$ is the estimated treated effect using propensity score weighting.

2.3.5. Conventional regression adjustment. For comparative purposes, we used regression adjustment to estimate the reduction in mortality due to statin therapy. Logistic regression was used to regress 3-year mortality on an indicator variable denoting receipt of a statin prescription at discharge and a vector of confounding factors. This was done in three different fashions. First, 3-year survival was regressed on exposure to statins at discharge and age, cardiogenic shock, systolic blood pressure, glucose, family history of heart disease, white blood count, creatinine levels, and respiratory rate. Second, backwards model selection was used to create

a parsimonious regression model to predict mortality. Statin exposure and all 24 variables listed in Table I were initially entered in the regression model. Statin exposure was forced to remain in all models. Variables were eliminated using backwards selection, with a P-value of 0.05 required for retaining the variable in the model. The odds ratio and associated 95 per cent confidence intervals for the reduction in mortality due to statin therapy were determined. Third, the complex regression model developed in Section 2.3.4 was used to estimate the reduction in 3-year mortality due to statin therapy. This method is described as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1 Z_i + \beta X_{ij}$$

Then $\exp(\beta_1)$ is the estimated treated effect using conventional logistic regression adjustment.

2.4. Estimation of treatment effect—absolute and relative treatment effects

The above methods estimated the adjusted relative odds of mortality for treated compared to untreated patients. However, they do not take advantage of the ability of propensity score methods to replicate the design of a randomized control trial—*conditional on all observed covariates*. RCTs allow for the direct calculation of both the absolute and relative reduction in mortality. In the following section, we describe propensity score methods that allow for the direct computation of the absolute and relative reduction in mortality due to statin therapy. In what follows, let Y_{0ij} denote the observed outcome for the ith untreated subject in the jth quintile, and Y_{1ij} denote the observed outcome for the ith treated subject in the jth quintile. Similarly, let Y_{0i} denote the observed outcome for the ith untreated subject in the matched sample, and let Y_{1i} denote the observed outcome for the ith treated subject in the matched sample. Let n_{0j} and n_{1j} denote the number of untreated and treated subjects in the jth quintile, respectively, while n_0 and n_1 denote the number of untreated and treated patients in the matched sample, respectively.

2.4.1. Stratifying on the quintiles of the propensity score. Patients were stratified into quintiles according to the predicted probability of receiving a statin prescription prior to hospital discharge. Within each quintile, the absolute and relative reduction in mortality associated with statin therapy were determined. Effect sizes were pooled across strata to obtain an overall absolute and relative reduction in mortality. This can be described mathematically as

$$\delta_j = \frac{1}{n_{0j}} \sum_{i=1}^{n_{0j}} Y_{0ij} - \frac{1}{n_{1j}} \sum_{i=1}^{n_{1j}} Y_{1ij}$$

$$\gamma_j = \frac{(1/n_{0j}) \sum_{i=1}^{n_{0j}} Y_{0ij}}{(1/n_{1j}) \sum_{i=1}^{n_{1j}} Y_{1ij}}$$

Then δ_j and γ_j denote the stratum-specific absolute and relative treatment effects. Then

$$\delta = \frac{1}{5} \sum_{j=1}^{5} \delta_j$$

and

$$\gamma = \frac{1}{5} \sum_{j=1}^{5} \gamma_j$$

denote the pooled absolute and relative treatment effects, respectively.

Stratifying on the quintiles of the propensity score can be expected to eliminate approximately 90 per cent of the bias due to imbalance in the observed covariates between treated and untreated patients [1]. However, there may remain residual imbalance between treated and untreated subjects within each stratum. Thus, within-stratum regression adjustment, using logistic regression was employed to determine the reduction in mortality associated with statin therapy, after adjusting for age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission. This allowed the calculation of both the absolute and relative reduction in mortality due to statin exposure. Within-stratum estimates were pooled to obtain overall treatment effects.

We describe our methods, using the terminology suggested by Lunceford and Davidian [21]. Let

$$m^{(j)}(Z,X,\beta^{(j)})$$

denote the stratum-specific logistic regression model for the jth stratum, in which mortality is regressed on treatment status (Z) and a vector of covariates (X). We let $\beta^{(j)}$ denote the stratum-specific vector of regression coefficients. Then

$$\delta_j = n_j^{-1}\sum_{i=1}^{n_j} m^{(j)}(1,X_i,\hat{\beta}^{(j)}) - m^{(j)}(0,X_i,\hat{\beta}^{(j)})$$

denotes the average treatment effect in stratum j, where the treatment effect is estimated for each subject in the jth stratum. Then the stratified estimate of treatment effect with within-stratum regression adjustment is given by

$$\hat{\delta} = K^{-1}\sum_{j=1}^{K}\hat{\delta}_j$$

In the current analysis, we use $K=5$ since we are stratifying on the quintiles of the propensity score. We used Bootstrap methods [25], with 1000 Bootstrap iterations to determine the variability of this estimate. This allowed for the calculation of both 95 per cent confidence intervals as well as for testing the statistical significance of the treatment effect. Using similar methods we were able to compute the relative benefit of statin treatment.

2.4.2. Matching on the propensity score. We created matched pairs consisting of one treated and one untreated patient. A greedy-matching algorithm was used to match subjects using calipers that were defined to have a maximum width of 0.2 standard deviations of the logit of the estimated propensity score [19]. The relative and absolute reduction in mortality was determined and its statistical significance was assessed using methods appropriate for paired binary data [26]. These estimates can be described mathematically as

$$\delta = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_{0i} - \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i}$$

$$\gamma = \frac{(1/n_0)\sum_{i=1}^{n_0} Y_{0i}}{(1/n_1)\sum_{i=1}^{n_1} Y_{1i}}$$

Then, δ and γ are the estimates of absolute and relative treatment effects obtained from the matched analysis, respectively.

2.4.3. Weighting using the propensity score. Propensity score weighting was used to obtain absolute and relative estimates of treatment effects. While, there are a number of estimators based on weighting, we only examined the one that Lunceford and Davidian describe as having a 'double-robustness' property, which can be obtained as [21]

$$\hat{\delta}_{\mathrm{DR}} = n^{-1}\sum_{i=1}^{n}\frac{Z_i Y_i - (Z_i - \hat{e}_i)m_1(X_i,\hat{\beta}_1)}{\hat{e}_i} - n^{-1}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i - (Z_i - \hat{e}_i)m_0(X_i,\hat{\beta}_0)}{1-\hat{e}_i}$$

where $\hat{e}_i$, Z_i, and Y_i denote the estimated propensity score, the treatment assignment indicator, and the outcome, respectively, for the ith subject. Here $m_z(X,\beta_z) = E(Y|Z=z,X)$ is the regression of the response on X in group z, $z=0,1$ (the reader is referred elsewhere for more details [21]). Lunceford and Davidian provide details on estimating the variance of the estimator, thus allowing for significance testing and the construction of confidence intervals [21]. This estimator is described as having a 'double-robustness' property because the estimator remains consistent if either the propensity score model is correctly specified or if the two regression models are correctly specified. The relative reduction in mortality using propensity score weighting was obtained using the ratio of the two terms in the above formula. The standard error and 95 per cent confidence interval was estimated using 1000 bootstrap samples.

3. RESULTS

3.1. Cohort characteristics

Detailed clinical data were available on 11 524 patients admitted with a diagnosis of AMI. Encrypted health card numbers were missing for four records, which were excluded since linkage to the RPDB was not possible without a valid health card number. A further 1137 patients died during the index hospitalization, resulting in a cohort of 10 383 patients discharged alive from an index hospitalization for AMI. Of these, a further 1279 patients were removed from the analyses due to missing data on important confounding factors (vital signs on admission and laboratory test results). This resulted in a final sample size of 9104 AMI survivors who were used for the development of the propensity score model. Overall, 3049 (33.5 per cent) patients received a statin prescription at discharge. The 3-year mortality rate was 14.2 and 25.3 per cent for those who did and did not receive a prescription for a statin at discharge, respectively. The unadjusted difference in mortality at 3-years was statistically significant ($P<0.0001$). The final propensity score model included 256 variables: 27 main effects and 229 two-way interactions. Following the derivation of the propensity score model, 95 untreated patients had a lower estimated propensity score than any treated patient, and were excluded from all subsequent analyses. Thus, 9009 patients were used in all subsequent analyses. The four thresholds determining the quintiles of the propensity score were: 0.147, 0.235, 0.332, and 0.586. Statin prescribing rates within the five quintiles were 9.3, 18.4, 28.3, 46.1, and 67.1 per cent, respectively.

The distribution of important confounding factors between treated and untreated patients is described in Table I. Patients who did not receive a statin prescription at discharge tended to be older, were more likely to be female, were more likely to have presented with acute CHF/pulmonary oedema, were more likely to have a history of cancer, chronic CHF, to have

a higher admission heart rate, to have a higher admission respiratory rate, to have higher white blood count, and to have elevated potassium and creatinine compared to patients who received a statin prescription at discharge. In summary, patients receiving statins tended to be younger and healthier than patients who did not receive a statin prescription at discharge. Standardized differences are reported in Table II. In the unmatched sample, 9 of the 24 covariates had a standardized difference that, in absolute value, exceeded 10 per cent. In particular, clinically meaningful differences in creatinine, angina, heart rate, gender, haemoglobin, respiratory rate, family history of heart disease, age, and history of hyperlipidaemia existed between treated and untreated patients. Thus, there is strong evidence that treatment status was confounded with factors that are prognostic of AMI mortality.

Figure 1 depicts a non-parametric density estimate of the distribution of the estimated propensity for treated and untreated patients separately. Each distribution is bi-modal. Patients who did not receive a statin prescription at discharge tended to have lower propensity scores. Since the patients matched on the propensity score tend to have the same distribution of observed covariates [2], this figure provides further evidence of an imbalance in measured covariates between treated and untreated patients, and that treatment assignment was confounded with observed covariates.

Table II. Standardized differences between treated and untreated patients.

			Stratified analysis				
Variable	Unmatched sample	Matched sample	Q1	Q2	Q3	Q4	Q5
Diabetes	−0.9	−0.2	9.2	−8.1	0.4	8.9	4.4
Renal disease	−1.9	0.0	−2.4	1.9	−3.9	2.9	2.0
Systolic BP	2.1	−0.4	11.9	−4.7	4.6	−5.0	−0.2
Glucose	−3.7	1.5	11.1	−11.8	0.9	2.5	3.5
Smoker	4.2	0.6	−5.4	0.3	4.5	3.2	−9.6
Cancer	−4.6	0.0	−2.1	6.0	0.0	0.3	−4.1
Diastolic BP	4.8	−0.2	5.6	−2.7	6.0	−6.3	−1.6
Acute shock	−4.8	0.0	−10.5	−0.1	−3.5	2.8	8.1
Acute CHF/pulmonary oedema	−5.8	−0.5	−7.8	−5.8	5.0	−4.0	8.9
Potassium	−6.2	−0.6	−10.8	−0.8	−5.8	2.7	−0.9
WBC	−6.7	0.5	−8.4	−2.4	3.9	−3.4	5.6
High BP	6.8	−2.6	1.2	−6.0	2.5	4.8	7.1
CVA/TIA	−8.1	−0.3	−1.5	−7.1	−5.0	1.1	6.7
Sodium	8.1	−0.8	−5.0	11.6	0.8	3.6	−6.6
Congestive heart failure	−8.2	−1.0	−8.5	−3.7	−2.5	−4.8	3.7
Creatinine	−10.0	1.5	−10.3	−2.6	−3.1	2.2	8.1
Angina	10.1	2.1	12.6	−2.2	−3.2	−1.2	7.2
Heart rate	−12.2	−0.2	2.1	−11.4	−0.2	14.2	−1.1
Female	−16.9	2.5	−1.8	−1.3	−6.3	7.7	−5.8
Haemoglobin	17.1	−3.6	4.6	6.3	8.1	−8.0	−3.4
Respiratory rate	−17.2	−1.3	−10.0	−0.4	−3.1	4.1	0.5
Family history	20.2	−1.3	5.8	4.7	0.5	−2.4	−0.7
Age	−36.1	−0.2	−19.7	1.7	−6.0	5.4	1.8
Hyperlipidaemia	87.5	−0.3	−11.6	−5.3	−1.3	30.0	−4.8

Note: Each cell is the per cent standardized difference between treated and untreated patients.

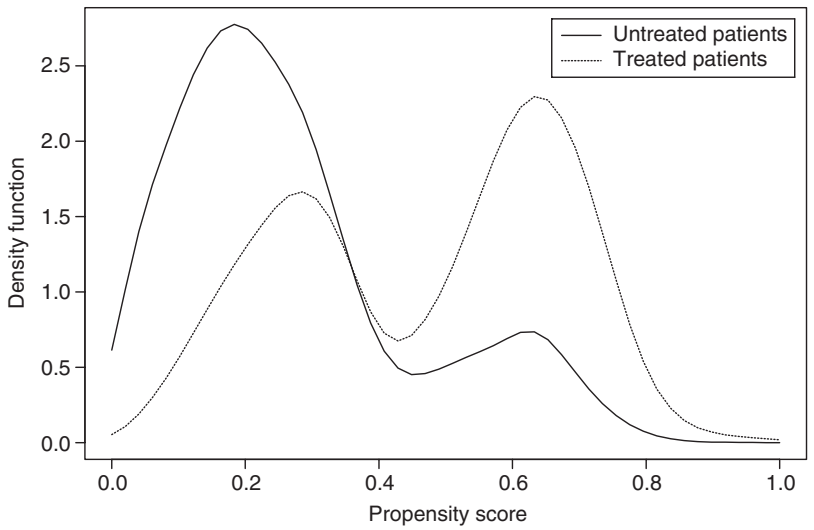

Figure 1. Distribution of propensity score in treated/untreated patients.

Due to the balancing properties of the propensity score, samples or stratum that are matched on the propensity score will tend to be balanced in their distribution of observed covariates. We now examine the degree to which either stratifying or matching on the propensity score resulted in the creation of either stratum or samples that are balanced in their distribution of the propensity score between treated and untreated subjects. Evidence of residual imbalance in the distribution of the propensity score would provide evidence that the distribution of the observed covariates are different between treated and untreated subjects in the same stratum or matched sample. Box plots of the estimated propensity score for treated and untreated patients within each quintile of the estimated propensity score are depicted in Figure 2. In general, the distribution of the propensity score appears to be similar between treated and non-treated patients within each quintile of the propensity score. One exception would seem to be in the fourth quartile, in which the median propensity score for treated patients is greater than that for untreated patients. Thus, using this graphical method of examination, stratifying on the quintiles of the estimated propensity score would seem to have created strata of patients who are similar in their propensity to receive treatment, and thus who are balanced in observed covariates. Figure 3 depicts five quantile–quantile plots that were used to compare the distribution of the propensity score for treated patients with that of untreated patients within each quintile of the propensity score. One can observe that in the second and third quintiles the distributions of the propensity score was nearly identical between treated and untreated patients. However, one can observe that in the first and fourth quintiles there is some evidence of residual imbalance. In the fifth quintile, the distribution of the propensity score amongst treated patients is slightly more positively skewed than that for untreated patients, however apart from this, the two distributions are very similar. The two-sample Kolmogorov–Smirnov test indicated that the distributions were comparable between treated and untreated patients in the second and third quintiles ($P = 0.1625$ and 0.0736, respectively), while there was evidence that the distributions were different between treated and untreated patients in the

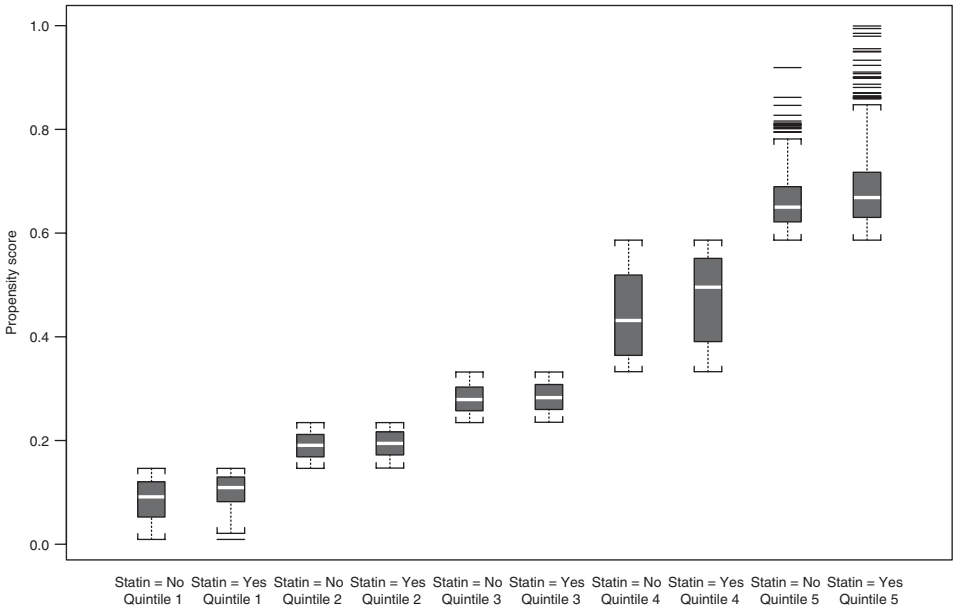

Figure 2. Comparison of propensity score for statin use. Figure 2 contains 10 boxplots. The shaded grey vertical box at the centre of each box plot denotes the middle 50 per cent of the data. Thus, the lower and upper ends of the box denote the 25th and 75th percentile, respectively. The solid white horizontal line through each shaded box denotes the median of the distribution. The vertical dotted lines (the 'whiskers') reach out to 1.5 $\times$ the interquartile range. Individual vertical lines beyond the whiskers denote individual extreme observations.

first, fourth and fifth quintiles ($P < 0.0001$ in all instances). Thus, by using either the quantile–quantile plots or the Kolmogorov–Smirnov test, there is some evidence of residual imbalance in the propensity score within some of the strata. Thus, there may be residual imbalance in observed characteristics between treated and untreated patients within stratum. Table III compares treated and untreated patients within each quintile of the propensity score for each patient characteristics. One can observe that overall treated and untreated patients are very similar within each quintile of the propensity score. There were only four exceptions to this. In the first quintile, untreated patients were, on average, 2 years older than treated patients. In the fourth quintile, untreated patients had a lower prevalence of history of hyperlipidaemia and tended to have a lower heart rate. Standardized differences for each covariate, and within each quintile of the propensity score are reported in Table II. There was evidence of residual imbalance, particularly in the first quintile.

Matching on the logit of the propensity score using calipers equal to 0.2 standard deviations of the logit of the propensity score resulted in the creation of 2364 matched pairs of treated and untreated patients. Thus, for 685 treated patients, no suitable control was found. This resulted in the elimination of 685 treated patients and 3691 untreated patients from the matched analysis. Figure 4 depicts non-parametric density estimates of the distribution of the propensity score within the matched sample for treated and untreated patients separately. One observes that matching on the propensity score resulted in a matched sample that is well

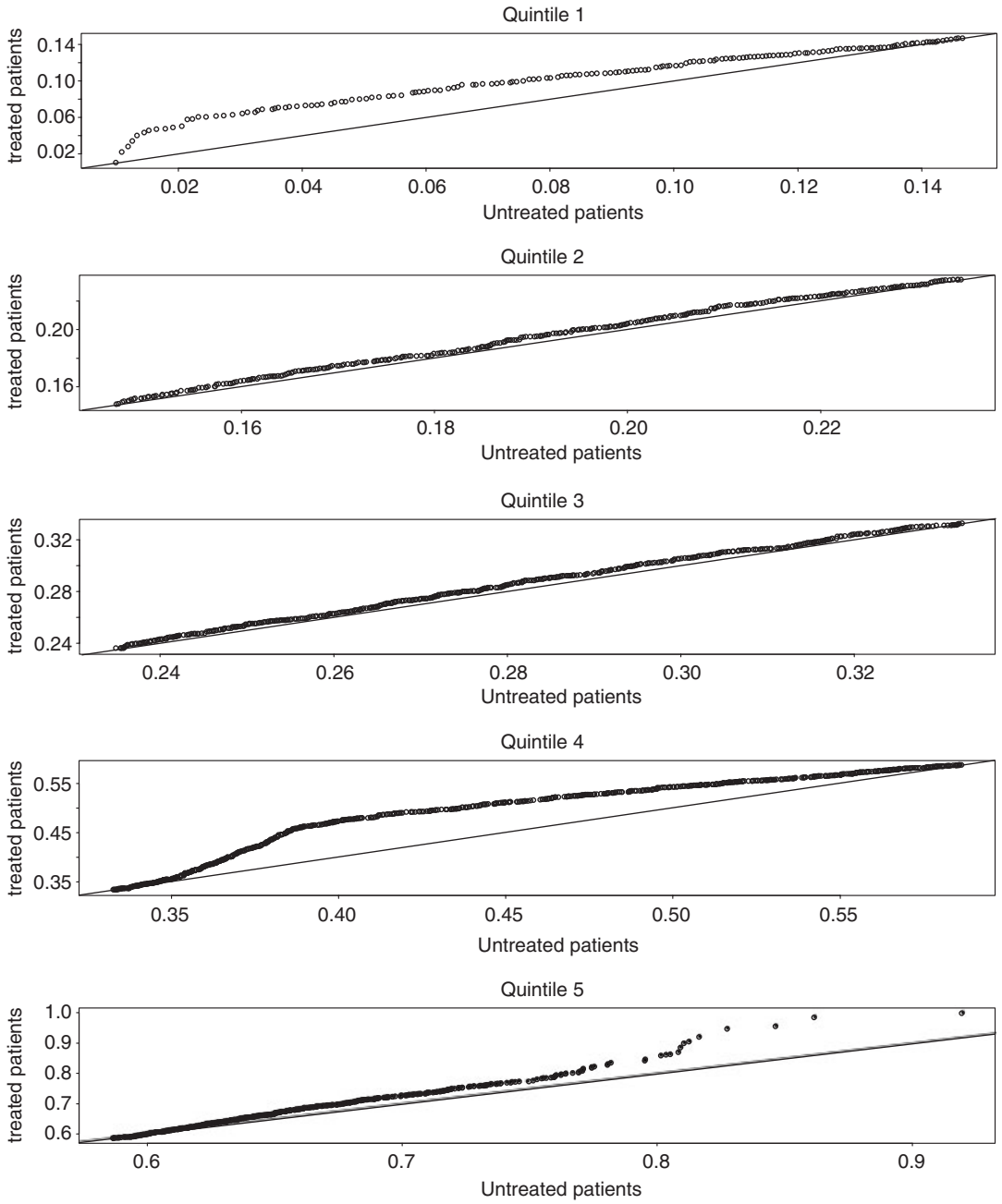

Figure 3. Comparison of distribution of PS between treated and untreated patients.

Table III. Comparison of treated and untreated patients within each quintile.

	Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5		
Variable	Statin: No $N = 1635$	Statin: Yes $N = 168$	P-Value	Statin: No $N = 1473$	Statin: Yes $N = 333$	P-Value	Statin: No $N = 1289$	Statin: Yes $N = 508$	P-Value	Statin: No $N = 970$	Statin: Yes $N = 830$	P-Value	Statin: No $N = 593$	Statin: Yes $N = 1210$	P-Value
Demographic															
Age	79.4±10.4	77.4±9.8	0.018	69.0±11.2	69.2±10.7	0.783	60.2±10.9	59.6±10.4	0.259	59.2±12.0	59.9±12.4	0.251	63.58±11.59	63.8±11.8	0.715
Female	871 (53.3%)	88 (52.4%)	0.826	553 (37.5%)	123 (36.9%)	0.837	329 (25.5%)	116 (22.8%)	0.234	270 (27.8%)	260 (31.3%)	0.105	162 (27.3%)	300 (24.8%)	0.249
Presenting characteristics															
Shock	28 (1.7%)	1 (0.6%)	0.273	9 (0.6%)	2 (0.6%)	0.982	5 (0.4%)	1 (0.2%)	0.527	3 (0.3%)	4 (0.5%)	0.557	0 (0.0%)	4 (0.3%)	0.161
Acute CHF	152 (9.3%)	12 (7.1%)	0.355	65 (4.4%)	11 (3.3%)	0.362	30 (2.3%)	16 (3.1%)	0.320	40 (4.1%)	28 (3.4%)	0.405	17 (2.9%)	55 (4.5%)	0.087
AMI risk factors															
Diabetes	494 (30.2%)	58 (34.5%)	0.248	548 (37.2%)	111 (33.3%)	0.185	138 (10.7%)	55 (10.8%)	0.941	209 (21.5%)	210 (25.3%)	0.060	155 (26.1%)	340 (28.1%)	0.381
CVA/TIA	321 (19.6%)	32 (19.0%)	0.856	120 (8.1%)	21 (6.3%)	0.258	41 (3.2%)	12 (2.4%)	0.356	49 (5.1%)	44 (5.3%)	0.811	51 (8.6%)	128 (10.6%)	0.187
Hyperlipidaemia	11 (0.7%)	0 (0.0%)	0.286	16 (1.1%)	2 (0.6%)	0.420	9 (0.7%)	3 (0.6%)	0.801	510 (52.6%)	557 (67.1%)	<.001	590 (99.5%)	1199 (99.1%)	0.359
High BP	740 (45.3%)	77 (45.8%)	0.887	690 (46.8%)	146 (43.8%)	0.322	434 (33.7%)	177 (34.8%)	0.636	492 (50.7%)	441 (53.1%)	0.308	279 (47.0%)	612 (50.6%)	0.159
Smoker	347 (21.2%)	32 (19.0%)	0.510	467 (31.7%)	106 (31.8%)	0.964	565 (43.8%)	234 (46.1%)	0.392	423 (43.6%)	375 (45.2%)	0.503	184 (31.0%)	323 (26.7%)	0.054
Comorbidities															
Angina	534 (32.7%)	65 (38.7%)	0.114	471 (32.0%)	103 (30.9%)	0.712	309 (24.0%)	115 (22.6%)	0.549	275 (28.4%)	231 (27.8%)	0.807	259 (43.7%)	572 (47.3%)	0.150
Cancer	75 (4.6%)	7 (4.2%)	0.803	53 (3.6%)	16 (4.8%)	0.300	28 (2.2%)	11 (2.2%)	0.993	16 (1.6%)	14 (1.7%)	0.951	16 (2.7%)	25 (2.1%)	0.398
CHF	144 (8.8%)	11 (6.5%)	0.320	59 (4.0%)	11 (3.3%)	0.549	16 (1.2%)	5 (1.0%)	0.648	27 (2.8%)	17 (2.0%)	0.314	19 (3.2%)	47 (3.9%)	0.470
Renal disease	13 (0.8%)	1 (0.6%)	0.779	3 (0.2%)	1 (0.3%)	0.735	1 (0.1%)	0 (0.0%)	0.530	2 (0.2%)	3 (0.4%)	0.533	3 (0.5%)	8 (0.7%)	0.691
Vital signs															
Systolic BP	144.8±34.1	148.9±34.4	0.141	151.5±31.6	150.0±32.2	0.437	148.8±30.0	150.1±29.0	0.385	152.9±29.3	151.4±29.5	0.286	147.55±28.70	147.5±29.6	0.967
Diastolic BP	79.8±19.5	80.9±20.2	0.484	83.8±18.2	83.3±18.2	0.661	86.2±18.0	87.3±18.15	0.254	87.1±18.3	86.0±18.5	0.186	83.40±15.94	83.1±17.0	0.748
Heart rate	93.5±27.0	94.0±26.8	0.797	86.2±23.8	83.6±22.4	0.065	78.1±19.1	78.1±18.8	0.976	75.4±19.4	78.2±20.2	0.003	83.69±24.60	83.4±25.0	0.829
Respiratory rate	23.6±7.4	22.9±6.9	0.228	20.8±5.1	20.7±4.7	0.947	19.7±3.7	19.6±3.6	0.560	19.9±4.1	20.1±4.3	0.381	20.21±4.76	20.2±5.1	0.929
Lab tests															
WBC	10.7±5.5	10.3±3.8	0.359	10.0±3.6	10.0±3.4	0.697	10.0±3.5	10.2±3.4	0.455	10.6±4.1	10.4±3.6	0.470	9.42±3.11	9.7±5.5	0.307
Haemoglobin	127.7±20.8	128.6±17.8	0.595	138.1±17.5	139.2±15.8	0.317	144.3±15.5	145.5±15.0	0.124	144.4±17.5	143.0±17.3	0.090	139.63±17.15	139.1±16.5	0.496
Sodium	138.0±4.6	137.8±4.7	0.531	139.0±3.4	139.3±3.1	0.065	139.8±3.1	139.8±2.9	0.882	139.2±3.4	139.3±3.3	0.453	139.35±4.44	139.1±3.2	0.161
Glucose	9.9±4.8	10.5±5.4	0.150	10.0±5.3	9.4±4.5	0.065	8.5±5.3	8.5±5.3	0.864	9.3±5.3	9.4±5.9	0.591	9.04±4.42	9.2±5.0	0.494
Potassium	4.2±0.6	4.1±0.5	0.211	4.1±0.6	4.1±0.5	0.893	4.0±0.5	4.0±0.5	0.271	4.0±0.5	4.1±0.5	0.574	4.11±0.56	4.1±0.5	0.851
Creatinine	119.8±75.2	113.1±53.8	0.258	99.5±45.0	98.4±43.1	0.667	93.3±35.0	92.2±30.7	0.559	96.5±42.8	97.4±41.5	0.644	99.20±39.39	103.4±61.3	0.130

Note: Continuous variables are reported as mean±standard deviation. Dichotomous variables are reported as number with condition (per cent).

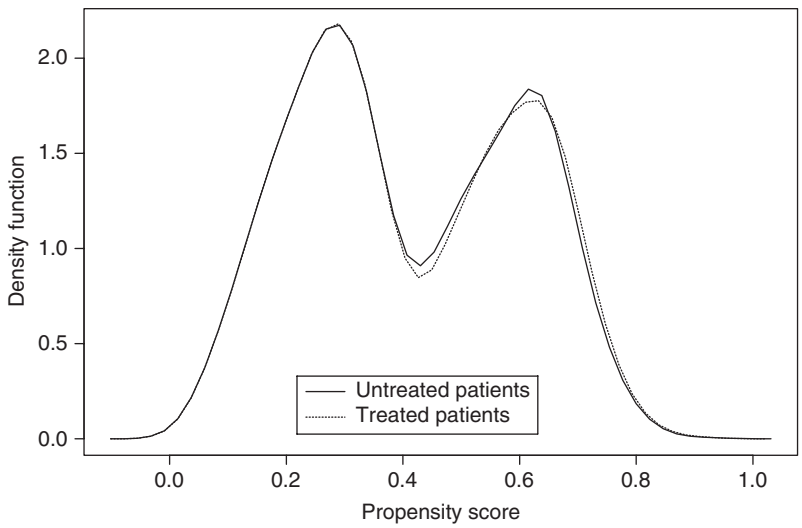

Figure 4. Distribution of propensity score in treated/untreated patients (matched analysis).

balanced in terms of the propensity score. This observation was confirmed by the two-sample Kolmogorov–Smirnov test, which did not find evidence that the two distributions of propensity scores were different from one another ($P = 0.6195$). Since the distribution of the propensity score was very similar between treated and untreated patients in the matched sample, one would expect the distribution of measured covariates to be very similar between treated and untreated patients in the matched sample. Table IV compares measured characteristics between treated and untreated patients who were matched on the propensity score. There were no significant differences in measured characteristics between treated and untreated patient in the matched sample. Standardized differences are reported in Table II. Of important note is that the largest standardized difference, in absolute value, was 3.6 per cent, indicating that very good balance has been achieved between treated and untreated patients in the matched sample.

3.2. Treatment effect—adjusted estimates

Adjusted estimates of the treatment effect of statin use on reducing 3-year mortality are described in Table V. The crude analysis suggested that the use of statins was associated with a 51 per cent reduction in the odds of death. However, as the above analyses demonstrate, treatment assignment is strongly confounded with factors that are prognostic of mortality, with younger, healthier patients being more likely to receive therapy than older, sicker patients. Stratifying on the quintiles of the propensity score led to the conclusion that statin therapy resulted in a 23 per cent reduction in the odds of death (odds ratio $= 0.77$, 95 per cent CI $= (0.66, 0.89)$). Within-quintile regression adjustment for age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission resulted in a negligible change in the estimated treatment effect (odds ratio $= 0.76$). Covariate adjustment assuming a linear relationship between the propensity score and the log-odds of mortality led one to conclude that statin

Table IV. Comparison of treated and untreated patients in matched sample.

Characteristic	Statin: No $N = 2348$	Statin: Yes $N = 2348$	P-Value
Demographic characteristics			
Age	63.33±12.55	63.30±12.53	0.951
Female	699 (29.8%)	726 (30.9%)	0.392
Presenting characteristics			
Shock	8 (0.3%)	8 (0.3%)	1.000
Acute CHF/pulmonary oedema	85 (3.6%)	83 (3.5%)	0.873
AMI risk factors			
Family history of CAD	919 (39.1%)	904 (38.5%)	0.642
Diabetes	579 (24.7%)	577 (24.6%)	0.944
CVA/TIA	158 (6.7%)	156 (6.6%)	0.906
Hyperlipidaemia	1071 (45.6%)	1068 (45.5%)	0.778
High BP	1107 (47.1%)	1077 (45.9%)	0.364
Current smoker	879 (37.4%)	886 (37.7%)	0.831
Comorbidities			
Angina	735 (31.3%)	758 (32.3%)	0.457
Cancer	58 (2.5%)	58 (2.5%)	1.000
CHF	68 (2.9%)	64 (2.7%)	0.728
Renal disease	7 (0.3%)	7 (0.3%)	1.000
Vital signs on admission			
Systolic BP	149.97±30.07	149.86±29.98	0.903
Diastolic BP	84.84±17.86	84.80±18.10	0.941
Heart rate	81.05±22.60	81.00±22.34	0.943
Respiratory rate	20.26±4.60	20.20±4.52	0.649
Laboratory values			
White blood count	10.12±3.66	10.14±3.94	0.870
Haemoglobin	141.70±17.54	141.08±16.92	0.203
Sodium	139.21±3.82	139.18±3.28	0.769
Glucose	9.18±4.98	9.26±5.43	0.608
Potassium	4.07±0.57	4.07±0.52	0.847
Creatinine	98.05±41.47	98.69±44.54	0.613

Note: Continuous variables are reported as mean±standard deviation, while dichotomous variables are reported as number with condition (per cent).

therapy resulted in a 16 per cent reduction in the odds of death at 3-years (odds ratio = 0.84, 95 per cent CI = (0.74, 0.96)). The estimated treatment effect was marginally magnified when one adjusted for age, cardiogenic shock at presentation, systolic blood pressure at admission, family history of heart disease, glucose level, white blood count, creatinine level, and respiratory rate at admission in addition to the propensity score (odds ratio = 0.80). When the propensity score was modelled as having either a quadratic relationship with mortality or when it was represented using a restricted cubic spline, one concluded that statin therapy resulted in a 19 per cent reduction in the odds of death (odds ratio = 0.81, 95 per cent CI = $(0.70, 0.93)$). In both instances, the models that incorporated a non-linear relationship between the propensity score and the log-odds of mortality fit the data significantly better than the simple model that only adjusted for the propensity score using a linear term ($P < 0.0001$). Matching on

Table V. Treatment effect for Statin at discharge—adjusted estimates.

Method	Odds ratio	95 per cent confidence interval	*P*-Value
Crude	0.49	(0.44, 0.55)	<0.0001
Propensity score methods			
Stratifying on PS quintiles	0.77	(0.66, 0.89)	0.0003
Stratifying on PS quintiles—within stratum regression adjustment	0.76	(0.65, 0.89)	0.0007
Covariate adjustment—linear term for PS	0.84	(0.74, 0.96)	0.0099
Covariate adjustment—quadratic terms for PS	0.81	(0.70, 0.93)	0.0033
Covariate adjustment—cubic spline for PS	0.81	(0.70, 0.93)	0.0028
Covariate adjustment—linear term for PS with adjustment for additional confounders	0.80	(0.69, 0.93)	0.0038
PS Matching	0.85	(0.72, 0.99)	0.0372
PS weighting—simple model	0.77	(0.64, 0.92)	0.0041
PS weighting—complex model	0.76	(0.63, 0.90)	0.0022
Direct regression adjustment			
Regression adjustment—simple predictive model	0.75	(0.65, 0.85)	<0.0001
Regression adjustment—backwards selection method	0.73	(0.64, 0.84)	<0.0001
Regression adjustment—complex predictive model	0.78	(0.67, 0.91)	0.0014

the propensity score resulted in the estimate that statin therapy reduced the odds of death by 15 per cent (odds ratio $= 0.85, 95$ per cent CI $= (0.72, 0.99)$). Finally, when propensity score weighting was used with a simple, pre-specified regression model, statin use was associated with a 23 per cent reduction in mortality (odds ratio $= 0.77$, 95 per cent CI $= (0.64, 0.92)$). This result was only marginally changed when a complex predictive model was constructed that included interactions and non-linear terms (odds ratio $= 0.76$).

For comparative purposes, when direct regression adjustment was used to estimate the treatment effect, statins were associated with a 25 per cent reduction in the odds of death (odds ratio $= 0.75, 95$ per cent CI $= (0.65, 0.85)$) after adjusting for age, shock, systolic blood pressure, glucose levels, family history of heart disease, white blood count, creatinine, and respiratory rate. When backwards selection was used, this result was modified negligibly (27 per cent reduction in the odds of death; odds ratio $= 0.73$; 95 per cent $= (0.64, 0.84)$). When the complex predictive model was used, the odds ratio was 0.78 (95 per cent CI $= (0.67, 0.91)$).

3.3. Treatment effects—direct estimation of absolute and relative reduction in mortality

Estimates of the absolute and relative reduction in mortality due to statin therapy are described in Table VI. Estimates of the absolute treatment effect ranged from 2.1 to 4.5 per cent, while

Table VI. Absolute and relative reduction in mortality due to statin treatment.

Method	Absolute risk reduction	95 per cent confidence interval	*P*-Value	Relative risk reduction	95 per cent confidence interval	*P*-Value
Matched analysis	2.1%	(0.1%, 4.0%)	0.0346	13.3%	(1.0%, 24.1%)	0.0350
Stratified analysis	4.5%	(2.4%, 6.6%)	<0.0001	17.0%	(5.7%, 25.6%)	0.0015
Stratified analysis—within stratum regression adjustment	3.7%	(1.5%, 5.7%)	0.0006	15.6%	(5.6%, 25.5%)	0.0034
Double-robust (weighted estimate)	3.4%	(1.1%, 5.7%)	0.0040	15.6%	(5.0%, 26.4%)	0.0083

estimates of the relative treatment effect ranged from a relative reduction in mortality of 13.3–17.0 per cent.

The stratified analysis resulted in an absolute reduction of 4.5 per cent. When within-stratum regression adjustment was used, the treatment effect was attenuated, with a 3.7 per cent absolute reduction in mortality being observed. This attenuation is an indication of the within-stratum residual confounding that remained after stratifying on the quintiles of the propensity score. As noted above, a small degree of within-stratum residual confounding was apparent in the first and fourth quintiles. The use of the weighted, or doubly robust estimator resulted in a further attenuation in both the estimated absolute and relative mortality reduction attributable to statin therapy. The use of propensity score weighting resulted in a 3.4 per cent absolute reduction in 3-year mortality and a 15.6 per cent relative reduction in mortality. The lowest absolute mortality reduction was obtained from the matched analysis. From the matched analysis one would infer that the use of statins resulted in a 2.1 per cent decline in 3-year mortality (13.3 per cent relative decrease in mortality). In all instances, both the absolute and relative benefits of statin therapy were statistically significant.

4. DISCUSSION

There is an increasing interest in the medical literature in the use of propensity score methods for estimating treatment effects using observational data. One of the advantages of propensity score methods is that one can explicitly determine the degree to which one has balanced measured characteristics between treated and untreated subjects. Our initial analysis demonstrated that patients who received a statin prescription were younger and healthier than those who did not, leading to treatment assignment and prognosis being confounded. Both stratifying and matching on the propensity score reduced the imbalance in measured characteristics between treated and untreated patients. However, greater balance was achieved through matching than through stratifying. After stratifying on the quintiles of the propensity score there was still minor residual imbalance within two of the quintiles. The greater balance achieved by matching was tempered by the reduced sample size, since there were many treated patients for whom no appropriate match could be found. Residual imbalance in the propensity score between treated and untreated patients within stratum or within a matched sample could indicate residual imbalance in measured covariates between treated and untreated patients. We explored

graphical methods to detect residual imbalance in stratum or matched samples. We explored within-stratum residual imbalance in the estimated propensity score using both box plots and quantile–quantile plots. Box plots have traditionally been used for examining the distribution of the propensity score between treated and untreated patients within strata of the propensity score. We demonstrated that the use of quantile–quantile plots might be more sensitive for detecting residual imbalance between treated and untreated patients.

The choice between matching and stratifying illustrates the classic tradeoff between variance and bias. Stratification may result in greater bias due to residual confounding within stratum. Matching may result in treated and untreated patients being discarded from the analysis, thus diminishing the precision of the estimated treatment effect. Both the absolute and relative reductions in mortality due to statin use were attenuated in the matched analysis relative to the stratified analysis. One explanation for this is the difference between the matched sample and the overall sample. The stratified analysis used the entire sample, whereas the matched analysis excluded those statin patients for whom a suitable control could not be found. It also resulted in the exclusion of the majority of the untreated patients, who on average were older and sicker than the treated patients. Thus, the matched analysis included patients who tended to be younger and healthier than did the stratified analysis. The differences between the matched sample and the overall cohort can be seen by comparing Tables I and IV. Similarly, Figures 1 and 4 highlight the differences in the populations for which treatment effects are being estimated. The differences between the stratified and matched analysis highlights the fact that they used different samples and address slightly different questions. The matched analysis estimates the treatment effect in patients who are similar to those who are treated, whereas the stratified analysis estimates the treatment effect amongst all patients who are eligible for treatment. The stratified analysis pools stratum-specific treatment effects, assigning equal weight to all strata. In particular, the stratum in which patients are least likely to receive therapy is assigned the same weight as the stratum consisting of those patients who have the greatest propensity to receive treatment. A limitation to the use of matching is that it requires a pool of potential controls or untreated patients at least as large as the number of treated patients. This was feasible in the current analysis, since only approximately a third of patients received a prescription for a statin at discharge. However, for other medications such as ASA or beta-blockers, which are dispensed to the majority of AMI survivors [27], the number of potential controls could be smaller than the number of treated patients.

In comparing results from stratified analyses with those from methods that use direct estimation, an important issue is frequently overlooked. Due to a phenomenon known as the non-collapsibility of the causal odds ratio, it is possible for the odds ratio for the entire cohort to be closer to one than any of the odds ratios for the strata [28–30]. This is frequently, though mistakenly, thought to be indicative of the presence of unmeasured confounding, but this need not be the case [28, 29]. This phenomenon may explain some of the divergent results that we observed. For instance, the odds ratio obtained by creating matched pairs using the propensity score was closer towards the null odds ratio than were the odds ratios obtained by stratifying on the propensity score. Thus, the non-collapsibilty of the odds ratio may explain, in part, the different estimates of treatment effects that we obtained.

Two recent systematic reviews of the use of propensity score methods in the medical literature have demonstrated that propensity score methods are often poorly implemented in the clinical literature [11, 12]. Four issues were apparent. First, little attention was paid to the construction of the propensity score. Frequently, the propensity score model was either

pre-specified or else obtained using automated variable selection methods such as backwards selection. Most authors assumed that the initial regression model was a satisfactory propensity score model. A structured approach similar to that described by Rosenbaum and Rubin was seldom employed. A second, related issue is that in only a minority of studies was an attempt made to determine the degree to which stratifying or matching on the propensity score balanced measured covariates between treated and untreated patients. One of the objectives of the current study was to carry out a detailed propensity score analysis, including an extensive exploration of residual confounding and the degree to which potential confounders were balanced between treated and untreated patients. Third, covariate adjustment using the propensity score was the most frequently used propensity score method. A limitation of covariate adjustment using the propensity score is that unbiased estimation of the treatment effect requires that the regression model has been correctly specified, while this is rarely examined in practice. In our study, we found that allowing for a quadratic relationship between the propensity score and the log-odds of death resulted in a qualitative change in the magnitude of the treatment effect. Our impression is that researchers using covariate adjustment with the propensity score rarely examine the fit of their model compared to more complex models. Furthermore, this method does not take advantage of the fact that creating strata that are matched on the propensity score allow one to mimic, conditional on measured covariates, the design of an RCT. Fourth, authors rarely computed absolute or relative treatment effects after matching or stratifying. Rather, authors tended to report adjusted estimates that were model-based (odds ratios or hazard ratios). In the current study, we presented both absolute and relative treatment effects as well as adjusted estimates obtained after stratifying or matching. Our intention was to highlight that propensity score methods can be used to obtain estimates of treatment effect that are in the same metric as those obtained from RCTs.

A recent meta-analysis has demonstrated that the use of statins in patients with coronary heart disease was associated with a 16 per cent reduction is the risk of all-cause mortality (risk ratio = 0.84, 95 per cent CI = (0.79, 0.89)) [15]. A second meta-analysis obtained similar treatment effects (all cause mortality was reduced by 15 per cent, 95 per cent CI = (8 per cent, 21 per cent)) [16]. The use of propensity score matching methods resulted in an identical estimate of the reduction in mortality due to statin therapy as was observed in a meta-analysis of statin RCTs. However, we do not want to use this result to suggest that matching is the optimal propensity score method. There are several reasons why one would not expect any of our results to be identical to those from meta-analyses of RCTs. First, our measure of exposure was whether the patient received a prescription for the given medication at hospital discharge. Research has shown that compliance with statin therapy is poor [31]. Second, patients who did not receive a prescription at hospital discharge may have received a prescription from their primary care physician or during a subsequent specialist referral. Third, although we adjusted for a long list of clinically important confounders derived from retrospective chart review, it is possible that there was unmeasured residual confounding. Thus, the reader is cautioned against using the meta-analysis as a gold standard against which to measure the performance of the competing propensity score methods examined in this paper. We would suggest that matching on the propensity score was optimal in our data set, not because, *a posteriori*, it resulted in an estimate that was closest to that obtained from meta-analyses, but because it was based on a matched sample in which there was virtually no imbalance in measured covariates between treated and untreated patients.

In conclusion, we compared the estimated reduction in mortality due to receiving a statin prescription at hospital discharge for an AMI using different propensity score methods on a single data set. We demonstrated the breadth of propensity score methods and that these methods allow the estimation of both adjusted as well as absolute and relative treatment effects. Earlier research has demonstrated that propensity score methods are often poorly implemented in the medical literature. Our intention, through carrying out an extended case study, was to demonstrate the breadth of the propensity score methods and how they can be more fully employed in clinical research.

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 in part by an operating grant from the Canadian Institutes of Health Research (CIHR) to the Canadian Cardiovascular Outcomes Research Team (CCORT). Dr Austin is supported in part by a New Investigator award from the CIHR. Dr Mamdani is supported by a New Investigator award from the New Emerging Teams (NETs) of the Canadian Institutes of Health Research (CIHR). The authors would like to thank two anonymous referees for their comments that improved the final manuscript.

REFERENCES

1. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
3. D'Agostino RB. Propensity score methods for bias reduction in the comparison of treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.
4. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* 1997; **127**:757–763.
5. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *American Journal of Epidemiology* 1999; **150**:327–333.
6. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Annals of Internal Medicine* 2002; **137**:693–695.
7. Wang J, Donnan PT. Propensity score methods in drug safety studies: practice, strengths and limitations. *Pharmacoepidemiology and Drug Safety* 2001; **10**:341–344.
8. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and Drug Safety* 2000; **9**:93–101.
9. Rosenbaum PR. *Observational Studies*. Springer: New York, NY, 1995.
10. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2001; **2**:169–188.
11. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods give similar results to traditional regression modelling in observational studies: a systematic review. *Journal of Clinical Epidemiology* 2005; **58**:550–559.
12. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modelling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* 2004; **13**:841–853.
13. Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM. *Quality of Cardiac Care in Ontario*. Institute for Clinical Evaluative Sciences: Toronto, Ontario, 2004.
14. Normand SLT, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following an acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**:387–398.
15. Wilt TJ, Bloomfield HE, MacDonald R, Nelson D, Rutks I, Ho M, Larsen G, McCall A, Pineros S, Sales A. Effectiveness of statin therapy in adults with coronary heart disease. *Archives of Internal Medicine* 2004; **164**:1427–1436.
16. Cheung BMY, Lauder IJ, Lau CP, Kumana CR. Meta-analysis of large randomized trials to evaluate the impact of statin on cardiovascular outcomes. *British Journal of Clinical Pharmacology* 2004; **57**:640–651.

17. Austin PC, Tu JV. Bootstrap methods for developing predictive models in cardiovascular research. *The American Statistician* 2004; **58**:131–137.
18. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
19. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya Series A* 1973; **35**:417–446.
20. Harrell Jr FE. *Regression Modeling Strategies*. Springer: New York, NY, 2001.
21. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
22. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
23. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician* 2004; **58**:272–279.
24. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**:387–394.
25. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: London, 1993.
26. Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine* 2004; **23**:65–75.
27. Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative *versus* clinical data. *Statistics in Medicine* 2005; **24**:1563–1578.
28. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**:761–768.
29. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science* 1999; **14**:29–46.
30. Rothman KJ, Greenland S. *Modern Epidemiology*. Lippincott Williams & Wilkins: Philadelphia, PA, 1998.
31. Jackevicius CA, Mamdani M, Tu JV. Adherence with statin therapy in elderly patients with and without acute coronary syndromes. *Journal of the American Medical Association* 2002; **288**:462–467.

