Comparing treatment effects after adjustment with multivariable Cox proportional hazards regression and propensity score methods

Edwin P. Martens MD, PhD1,2, Anthonius de Boer MD, PhD1, Wiebe R. Pestman PhD2, Svetlana V. Belitser MD, PhD1, Bruno H. Ch. Stricker MD, PhD3 and Olaf H. Klungel PharmD, PhD1*

1Department of Pharmacoepidemiology & Pharmacotherapy, Utrecht Institute of Pharmaceutical Sciences (UIPS), Utrecht University, Utrecht, The Netherlands
2Centre for Biostatistics, Utrecht University, Utrecht, The Netherlands
3Department of Epidemiology & Biostatistics, Erasmus MC, Rotterdam, The Netherlands

SUMMARY

Purpose To compare adjusted effects of drug treatment for hypertension on the risk of stroke from propensity score (PS) methods with a multivariable Cox proportional hazards (Cox PH) regression in an observational study with censored data.

Methods From two prospective population-based cohort studies in The Netherlands a selection of subjects was used who either received drug treatment for hypertension (n = 1293) or were untreated ‘candidates’ for treatment (n = 954). A multivariable Cox PH was performed on the risk of stroke using eight covariates along with three PS methods.

Results In multivariable Cox PH regression the adjusted hazard ratio (HR) for treatment was 0.64 (CI 95%: 0.42, 0.98). After stratification on the PS the HR was 0.58 (CI 95%: 0.38, 0.89). Matching on the PS yielded a HR of 0.49 (CI 95%: 0.27, 0.88), whereas adjustment with a continuous PS gave similar results as Cox regression. When more covariates were added (not possible in multivariable Cox model) a similar reduction in HR was reached by all PS methods. The inclusion of a simulated balanced covariate gave largest changes in HR using the multivariable Cox model and matching on the PS.

Conclusions In PS methods in general a larger number of confounders can be used. In this data set matching on the PS is sensitive to small changes in the model, probably because of the small number of events. Stratification, and covariate adjustment, were less sensitive to the inclusion of a non-confounder than multivariable Cox PH regression. Attention should be paid to PS model building and balance checking. Copyright © 2007 John Wiley & Sons, Ltd.

INTRODUCTION

Cox proportional hazards regression (Cox PH) has been widely used as an adjustment technique in observational studies with censored data. Often there is one variable of interest (the ‘treatment’ effect) and a set of covariates (confounders) that are used as independent variables to explain a dichotomous outcome variable. When these covariates are included in
the model it can be said that the treatment effect is adjusted for the influence of the observed confounders. An alternative approach in such cases is to use the propensity score (PS), a method originally proposed by Rosenbaum and Rubin in 1983. With this approach the focus is on the imbalance of covariates between treatment groups, which can be seen as a result of the non-random assignment of treatments to patients. Therefore, in the PS method first attention is directed to balance treatment groups with respect to the observed covariates and second to estimate the treatment. In fact, a randomized controlled trial (RCT) has a similar two-step procedure: first balancing treatment groups and second estimating treatment effect. Of course, a randomization procedure aims at balancing treatment groups on all confounders, where the PS can only handle confounders that are observed.

This approach is theoretically different from a Cox PH, linear or logistic regression model where an adjusted treatment effect is estimated by using the observed covariates as additional explanations for the variation in the response variable. This means that the method of PSs is an alternative for model-based methods as far as estimation of a treatment effect is concerned; it is no alternative when the objective is to model and estimate the influence of the observed confounders on the response variable.

The PS is defined as the conditional probability of being treated given the values of covariates. In general this probability is unknown but can be estimated using logistic, probit or discriminant analysis, where treatment is considered the dependent variable. It has been shown that a treated patient and an untreated control with the same PS or classes of subjects with the same PS tend to have the same distribution of covariates. This means that the PS can be used as a single matching or stratification variable to reduce confounding due to observed covariates. Furthermore, the distribution of the PS can be compared between treatment groups, revealing for which part of the treated patients no controls are available and vice versa. This possible lack of overlap is essential information when treatment groups are to be compared on some response variable, something that is seldom done or reported when a Cox PH, linear or logistic regression analysis has been performed.

PS methods to adjust for confounding are increasingly used in the medical literature, different PS methods and model-based adjustment techniques have been less frequently compared. In a recent simulation study, stratification on the PS was compared to logistic regression analysis and in some other studies a PS analysis was performed together with a model-based adjustment technique (among others). Our study objective was to systematically compare the effect of drug treatment for hypertension on the risk of stroke between a multivariable Cox PH regression and three PS methods.

**MATERIALS AND METHODS**

**Data**

The data we used have been described by Klungel et al. and come from two prospective population-based cohort studies in The Netherlands. Briefly, the first study, the Monitoring Project on Cardiovascular Risk Factors, was conducted from 1987 through 1991 as a cross-sectional study in Amsterdam, Maastricht and Doetinchem (62% agreed to participate). In Doetinchem, subjects were followed up through general practice records. The second study, the Rotterdam Study, was started in 1990 in Rotterdam as a population-based prospective follow-up study. All residents of a suburb of Rotterdam aged 55 years or older were invited to participate (78% agreed). The baseline measurements continued until 1993. In total 1293 treated hypertensives and 954 untreated ‘candidates’ for treatment were used for analysis, where the incidence of stroke was the response. The overall incidence rate was 4.2%, with 42 cases in the treated and 53 cases in the untreated patients. The selection of untreated controls was based on high blood pressure and the existence of other common cardiovascular risk factors. The following confounding factors were available for analysis: history of cerebrovascular disease (cerebrovascular accident, CVA), age, sex, diabetes, total cholesterol, body mass index, smoking, previous cardiovascular disease (CVD), previous myocardial infarction (MI), previous transient ischemic attack (TIA), family history of MI and HDL-cholesterol.

Three sets of covariates were defined. The first set, motivated by Klungel et al., consists of a selection of eight covariates (history of CVA, age, sex, diabetes, total cholesterol, body mass index, smoking and previous CVD). The second set consists of all available covariates. In order to investigate the sensitivity of the estimated treatment effect for the inclusion of a non-confounder, we created a third set of covariates. This simulated binary non-confounder was not correlated with treatment (equally balanced over treatment groups) nor with all other covariates in the model, but strongly associated with outcome (the incidence of stroke). Inclusion of such a risk factor will not change the estimated treatment effect in linear models, but it will change the effect in models
like logistic regression or Cox PH regression. By including this non-confounder we are able to compare the sensitivity to the results of the various methods.

**Multivariable Cox proportional hazards regression**

We used a multivariable Cox PH regression to model the time and the incidence of stroke (see for instance Therneau, SPSS 11.0). By adding the covariates to the model adjustment for confounding is achieved and an adjusted treatment effect is estimated. As the number of events per covariate was too low to use all covariates with this method, only the first and third set of covariates were used; a maximum of 10 events per covariate is advised in the literature.

**Propensity score methods**

**Achieving balance.** With treatment as the dependent and the three different sets as covariates we used logistic regression analysis to estimate the PS (SPSS 11.0). Some interactions and higher-order terms were added in order to improve the balance. In this model the number of ‘events’ (i.e. the lower of the number of treated and untreated patients) was sufficient to include these extra terms, in contrast to the multivariable Cox PH regression were the number of events (i.e. the number of strokes) is rather limited. Even when overfitting takes place in the PS model by a large number of terms, this is not of great concern, because it is not the intention to make inferential statements concerning the relationship between treatment and covariates. Instead we will focus on the balance of covariates between groups that will result when PS methods are used.

For a similar reason, we did not check goodness-of-fit (GOF) or the discrimination of the PS model (as is frequently done by reporting the Hosmer–Lemeshow GOF or the area under the receiver operator characteristic curve or c-statistic): the issue is not to predict treatment or to estimate coefficients. By adding interactions and higher-order terms to the PS model we selected only potential confounding factors, that is, those terms that showed at least a moderate relationship with the response. By this strategy, we clearly express that we focus on the problem of confounding and not on making the best predictive model for treatment. On the other hand, inclusion of some other terms or misspecification of the model does not seem to be of major concern.

**Checking balance on covariates.** A check on the balance on covariates achieved by the PS is essential for this method, although not always done or reported in the literature. To perform this check we used a stratified logistic regression analysis with treatment as the dependent, covariates as independents (LogXact 2.1) and with strata based on the quintiles of the PS (strata referred to as ‘fifths’). We also applied the standard method where for every covariate and every stratum of the PS the difference between treatment groups is assessed and tested. We prefer the stratified multivariable method because many separate comparisons, having reduced power within strata, will then be avoided. Another reason is that balance should be checked conditional on other covariates, which can be achieved when using a stratified multivariable check. Ideally, within subclasses of the PS all covariates should be balanced and differences between treatment groups should disappear.

**Estimating adjusted treatment effects.** We estimated an adjusted treatment effect in three ways: (1) stratification on the PS, (2) matching on the PS and (3) using the PS as a covariate.

1. Stratification on the PS was based on its quintiles. The resulting categorical variable was used in a Cox PH regression with stroke as the dependent and treatment as the only independent (S-Plus 6.2). The interaction between treatment and PS was tested in order to compare differences in treatment effect within strata.

2. Matching on the PS was based on pair-matching. This means that for every treated subject only one control was selected. A greedy algorithm was used (SAS 8.0) and resulted in such pairs of subjects by randomly selecting a case and matching this to the control with the smallest difference in PS. This process was continued until no more controls could be found that differed less than 0.1 in PS.

3. The third estimation method is to use the PS as a continuous covariate in a Cox PH regression replacing all single covariates. Although this method has often been used in practice, it is not recommended because too much weight is given to the absolute value of estimation of the PS. Another reason is that assumptions have to be made about the functional relationship between the PS and the response.

These three sets of covariates are combined with the four different adjustment methods, as is summarized in Table 1.
RESULTS

Actual imbalance on covariates

In Table 2 the means or percentages of all covariates for both treatment groups are given, including the univariate test result on their differences. Most of the odds ratios (ORs) are not close to one indicating imbalance on these covariates between groups. The covariates diabetes, family history of MI and total cholesterol are reasonably well balanced between treatment groups, whereas sex, previous TIA and previous CVD are clearly imbalanced between treatment groups.

Balance creating properties of the propensity score

As a first impression of the balance created by the PS we investigated the overlap in PS distributions between both treatment groups using the first set of covariates (Figure 1). As could be expected, the untreated group tends to have lower scores: 18% of the untreated patients compared to only 3% of the treated patients have a probability of being treated of less than 0.30, whereas PSs higher than 0.70 are found for 13% of the untreated and for more than 35% of the treated patients. On the other hand there is considerable overlap: only 1.2% of the subjects have a PS outside the range of the other group.

For a further check on the balance of the covariates and some interactions we used a stratified logistic regression with treatment as the dependent variable. The results are given in Table 3. Most of the ORs are near 1 and none reached a significance level of 0.10. The relatively low OR of 0.57 for sex (with very large confidence interval (CI)) is mainly due to the inclusion of two interaction terms with sex, giving this coefficient a less straightforward interpretation (in a model without these interactions the OR for sex is 0.98).

The check for balance for sex is given in Table 4. All ORs within the five strata of the PS are non-significant and closer to 1 than the highly significant OR for the total sample.

First set of covariates

The estimated hazard ratio (HR) in the Cox PH regression adjusted for the eight covariates was 0.64 with 95% confidence interval (CI95%) from 0.42 to

<table>
<thead>
<tr>
<th>Method of analysis</th>
<th>Set 1: 8 covariates</th>
<th>Set 2: 12 covariates</th>
<th>Set 3: set 1 plus balanced covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multivariable Cox PH regression</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Cox PH regression, stratification on PS</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Cox PH regression, matching on PS</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Cox PH regression, PS as covariate</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Cox PH, Cox proportional hazards; PS, propensity score.
*Analysed.
1Not analysed because of small number of events per covariate.

Table 2. Means or percentages of covariates for treated and untreated candidates for treatment, odds ratios and univariate significance tests

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Treated (n = 1293)</th>
<th>Untreated (n = 954)</th>
<th>Odds ratio</th>
<th>95%CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of CVA (%)</td>
<td>7.7</td>
<td>5.1</td>
<td>1.53</td>
<td>1.08, 2.18*</td>
</tr>
<tr>
<td>Sex (% men)</td>
<td>34.3</td>
<td>47.5</td>
<td>0.58</td>
<td>0.49, 0.69*</td>
</tr>
<tr>
<td>Smoking (%)</td>
<td>21.0</td>
<td>24.1</td>
<td>0.84</td>
<td>0.69, 1.02</td>
</tr>
<tr>
<td>Diabetes (%)</td>
<td>7.3</td>
<td>7.0</td>
<td>1.05</td>
<td>0.76, 1.45</td>
</tr>
<tr>
<td>Previous CVD (%)</td>
<td>24.1</td>
<td>17.1</td>
<td>1.54</td>
<td>1.24, 1.90*</td>
</tr>
<tr>
<td>Previous MI (%)</td>
<td>11.4</td>
<td>9.2</td>
<td>1.26</td>
<td>0.96, 1.67</td>
</tr>
<tr>
<td>Previous TIA (%)</td>
<td>4.4</td>
<td>2.5</td>
<td>1.79</td>
<td>1.10, 2.90*</td>
</tr>
<tr>
<td>Family history of MI (%)</td>
<td>10.5</td>
<td>9.4</td>
<td>1.14</td>
<td>0.87, 1.51</td>
</tr>
<tr>
<td>Age</td>
<td>65.1</td>
<td>65.8</td>
<td>1.00</td>
<td>0.99, 1.00</td>
</tr>
<tr>
<td>Body mass index</td>
<td>27.8</td>
<td>26.8</td>
<td>1.07</td>
<td>1.05, 1.09*</td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>6.56</td>
<td>6.61</td>
<td>0.97</td>
<td>0.90, 1.04</td>
</tr>
<tr>
<td>HDL-cholesterol</td>
<td>1.26</td>
<td>1.32</td>
<td>0.62</td>
<td>0.49, 0.79*</td>
</tr>
</tbody>
</table>

CVA, cerebrovascular accident; CVD, cardiovascular disease; MI, myocard infarction; TIA, transient ischemic attack.
*Different from an odds ratio of 1 at a significance level of 0.05, two-sided, using likelihood ratio test.

When stratification on the PS was used a slightly smaller HR was found (0.58 vs. 0.64), indicating a slightly larger treatment effect, estimated somewhat more precisely (CI95%: 0.38, 0.89). The treatment effects within the five strata did not differ significantly from each other ($p = 0.89$). Matching on the PS leads to an even larger treatment effect (0.49), somewhat less precisely estimated mainly because of a reduced number of observations in the analysis. Using the PS as a covariate gives similar results as the multivariable Cox PH regression.

**Second set of covariates**

In the second set, an adjusted treatment effect is estimated for the three different PS methods when four covariates were added to the first set. Because of the low number of events per covariate a multivariable Cox PH regression was not performed. For all PS methods we found a similar downward shift in the hazard ratio of around 7% compared to the first set of covariates, as well as a smaller confidence interval (Table 6).

**Third set of covariates: first set plus balanced covariate**

In the third set a balanced covariate was added to the first set to check the sensitivity of the various methods.
Table 5. Unadjusted treatment effects and adjusted effects with the first set of covariates using multivariable Cox PH, stratification on the PS, matching on the PS and PS as covariate

<table>
<thead>
<tr>
<th>Method of analysis</th>
<th>HR</th>
<th>95% CI</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>0.54</td>
<td>0.36, 0.82</td>
<td>2246</td>
</tr>
<tr>
<td>Multivariable Cox PH regression</td>
<td>0.64</td>
<td>0.42, 0.98</td>
<td>2134</td>
</tr>
<tr>
<td>Cox PH regression, stratification on PS</td>
<td>0.58</td>
<td>0.38, 0.89</td>
<td>2136</td>
</tr>
<tr>
<td>Cox PH regression, matching on PS</td>
<td>0.49</td>
<td>0.27, 0.88</td>
<td>1490</td>
</tr>
<tr>
<td>Cox PH regression, PS as covariate</td>
<td>0.64</td>
<td>0.41, 0.99</td>
<td>2134</td>
</tr>
</tbody>
</table>

Cox PH, Cox proportional hazard; HR, hazard ratio; PS, propensity score; n, number of observations.

Table 6. Adjusted treatment effects with the second set of covariates using stratification on the PS, matching on the PS and PS as covariate

<table>
<thead>
<tr>
<th>Method of analysis</th>
<th>HR</th>
<th>95% CI</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox PH regression, stratification on PS</td>
<td>0.53</td>
<td>0.35, 0.83</td>
<td>2037</td>
</tr>
<tr>
<td>Cox PH regression, matching on PS</td>
<td>0.45</td>
<td>0.25, 0.84</td>
<td>1488</td>
</tr>
<tr>
<td>Cox PH regression, PS as covariate</td>
<td>0.57</td>
<td>0.36, 0.89</td>
<td>2122</td>
</tr>
</tbody>
</table>

Cox PH, Cox proportional hazard; HR, hazard ratio; PS, propensity score; n, number of observations.

DISCUSSION

Three PS methods were compared with a multivariable Cox PH regression to estimate an adjusted effect of drug treatment for hypertension on the incidence of stroke. Matching and stratification on the PS gave a somewhat larger treatment effect than when a multivariable Cox PH regression was used or when the PS was used as covariate. PS methods had the possibility to include more covariates (not performed in the multivariable Cox model), which gave a similar shift in the treatment effect in all PS methods. When a balanced covariate was added, the smallest change was found by stratification on the PS and when the PS was used as covariate; when the multivariable Cox PH regression or matching on the PS was used the change was large.

We contributed to the application of PS methods in medical science by giving a systematic comparison between these methods and a model-based adjustment approach in a real life dataset with many covariates and a relatively low number of events. Furthermore, we pointed at the difficulties in finding the best PS model and in checking the balance between treatment groups. We also tested the sensitivity of the models against the addition of more covariates, including a balanced one.

In the medical literature application of PS methods is becoming more widespread. In most studies only one of the methods has been used, whereas only some compare the results with a model-based approach. Because in most of these studies the same set of covariates was used in the PS together with covariate adjustment, the conclusion that ‘no differences were found when a PS method was used’, is not surprising. Often it is unclear how the model was created and whether the balance was sufficient.14

A recent systematic comparison of a PS method and multivariable logistic regression analysis with a low number of events can be found in Cepeda et al.7 In a
simulated data set the number of confounding variables, the strength of associations, the number of events and the strength of the exposure were varied. It was concluded that the estimation of the treatment effect by means of PSs was less biased, more robust and more precise than logistic regression when there were seven or fewer events per confounder. With more than seven events logistic regression analysis was recommended. Unfortunately they used only the known PS model, the one that was used for generating the data, so that the step of reaching balance could be skipped. Furthermore, they used the PS only as a categorical variable in the final analysis, where covariate adjustment or matching on the PS could have been used.

Our study has some limitations. First, the data set used was already to some extent balanced by choosing a control group that consisted of untreated candidates for treatment. A more general control group would produce less overlap in the distributions of covariates and could lead to larger differences between the methods. On the other hand, the more comparable the groups are, the more the differences in treatment effect can be contributed to the methods instead of the specific data set used. A second limitation is that only greedy pair-matching was used. Unfortunately, a more optimal method could not be used because no large pool of controls was available. To use five instead of another number of classes goes back to Cochran, who stated that a 90% bias reduction is expected when stratifying was based on the quintiles. Also 7 and 10 strata were used, but this did not change the main results.

Furthermore, one can comment that the multivariable way of checking balance will leave questions whether this balance is sufficient and whether imbalances within strata exist. It can be shown that the balance on all these observed covariates is even better than could be expected in a RCT. In a RCT it is expected that on average 1 in 10 of the terms is significant at the 0.10 levels, where in our model none was found. Of course, randomization takes care of all covariates, also the unobserved ones. We also checked the balance on all of the eight covariates separately within the five strata of the PS. We found that only 1 out of 40 comparisons turned out to be significant at the 0.10 level, where 4 are to be expected in case of randomization.

A last comment concerns the precision of the differences found between the different methods. No confidence intervals are given for these differences, so that it is unclear to what extent the results are sensitive for the specific sample.

Application of PS methods is not a straightforward task. There exist some practical difficulties in applying this intuitively appealing method. The first is the check for balance; a crucial step after a PS model has been made. There are no general rules available for the practical user how this check needs to be performed. We used a multivariable stratified analysis, but it remains unclear whether this is the best way to check balance. Another difficulty is when to stop adding interactions and higher-order terms to the PS model when an acceptable balance has not yet been reached. There is hardly any limit in the number of terms that can be added, because estimating coefficients is not an objective. Therefore, measures of GOF, area under the ROC and predictability of the model should not be used as a guideline. The PS model is meant to adjust for confounding, which means that terms are to be considered that have a relationship with treatment as well as the response. The relationship with treatment can be checked in the PS model itself (as usual in logistic regression analysis), but the relationship with the response should come from outside this model. In general only terms should be included in the PS model that have an empirical or logical relationship with the response, because otherwise effort is wasted in attempting to balance non-confounders. This contradicts the idea that the same PS model can be used for different response variables.

Concerning the sensitivity of the inclusion of a non-confounder, stratification on the PS (and covariate adjustment) performed better than matching and multivariable Cox PH. Matching on the PS seems also to be a rather sensitive method when there is a small number of events, like in our data set. It is recommended to perform PS methods, but special attention should be given to the PS model building and balance checking phases.

**KEY POINTS**

- At least as a check for consistency PS methods should be applied whenever possible.
- PS methods give the possibility of including more covariates which is interesting when the number of events is rather small.
- Less attention should be paid to GOF measures of the PS model and more attention should be paid to check the balancing properties.
- It is not clear which PS method is best, but stratification was less sensitive to inclusion of non-confounders.
ACKNOWLEDGEMENTS
We thank W.M. Monique Verschuren from the Centre for Prevention and Health Services Research, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands and Monique M.B. Breteler from the Department of Epidemiology & Biostatistics, Erasmus MC, Rotterdam, The Netherlands.

REFERENCES