2014

The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments

Peter C Austin, *Institute for Clinical Evaluative Sciences*

Statistics
in Medicine

Tutorial in Biostatistics

Received 23 October 2012, Accepted 3 September 2013 Published online 30 September 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5984

The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments

Peter C. Austin[a,b,c*†]

Propensity score methods are increasingly being used to estimate causal treatment effects in observational studies. In medical and epidemiological studies, outcomes are frequently time-to-event in nature. Propensity-score methods are often applied incorrectly when estimating the effect of treatment on time-to-event outcomes. This article describes how two different propensity score methods (matching and inverse probability of treatment weighting) can be used to estimate the measures of effect that are frequently reported in randomized controlled trials: (i) marginal survival curves, which describe survival in the population if all subjects were treated or if all subjects were untreated; and (ii) marginal hazard ratios. The use of these propensity score methods allows one to replicate the measures of effect that are commonly reported in randomized controlled trials with time-to-event outcomes: both absolute and relative reductions in the probability of an event occurring can be determined. We also provide guidance on variable selection for the propensity score model, highlight methods for assessing the balance of baseline covariates between treated and untreated subjects, and describe the implementation of a sensitivity analysis to assess the effect of unmeasured confounding variables on the estimated treatment effect when outcomes are time-to-event in nature. The methods in the paper are illustrated by estimating the effect of discharge statin prescribing on the risk of death in a sample of patients hospitalized with acute myocardial infarction. In this tutorial article, we describe and illustrate all the steps necessary to conduct a comprehensive analysis of the effect of treatment on time-to-event outcomes. © 2013 The authors. Statistics in Medicine published by John Wiley & Sons, Ltd.

Keywords: propensity score; observational study; propensity score matching; inverse probability of treatment weighting; survival analysis; event history analysis; confounding; marginal effects

1. Introduction

The Consolidated Standards of Reporting Trials (CONSORT) statement provides recommendations for the reporting of RCTs [1]. Its intent was to alleviate problems arising from inadequate reporting of RCTs. Among its recommendations is that, for RCTs with dichotomous outcomes, both relative and absolute measures of treatment effect be reported. The importance of absolute measures of treatment effect to clinical decision making has been described by a variety of authors [2–5]. In particular, reporting absolute reductions in the risk of an event allows one to report the number needed to treat, an important estimate for medical decision making [3, 4].

Time-to-event outcomes occur frequently in reports of RCTs in the medical literature [6]. In keeping with the intent of the CONSORT statement, a thorough analysis of the effect of treatment on survival in

[a]*Institute for Clinical Evaluative Sciences, Toronto, Canada*
[b]*Institute of Health Management, Policy and Evaluation, University of Toronto, Toronto, Canada*
[c]*Dalla Lana School of Public Health, University of Toronto, Toronto, Canada*
**Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences G1 06, 2075 Bayview Avenue Toronto, Ontario M4N 3M5, Canada.*
†*E-mail: peter.austin@ices.on.ca*

an RCT should include estimation of both relative and absolute measures of treatment effect. Because randomization ensures that, on average, treatment assignment is unconfounded with baseline covariates, outcomes can be compared directly between treatment groups. Thus, Kaplan–Meier estimates of survival functions in different treatment groups are often estimated. From these, the absolute reduction in the probability of an event occurring within a specified duration of follow-up can be determined, as can the number of patients needed to be treated to avoid one event occurring within a specified duration of follow-up. Similarly, by using a Cox proportional hazards model, the investigator can obtain an unbiased estimate of the relative change in the hazard of the event occurring because of the treatment.

Observational studies are increasingly being used to estimate the effects of treatments, exposures, and interventions on outcomes. In observational studies, there are often systematic differences in the distribution of baseline characteristics between treated and untreated subjects. Because of this, outcomes cannot be compared directly between treatment groups. Instead, statistical methods must be used to minimize the effects of confounding and obtain an unbiased estimate of treatment effect. Dorn has suggested that when designing an observational study, one should ask 'How would the study be conducted if it were possible to do it by controlled experimentation?' [7]. Thus, the design of an observational study should mimic that of a randomized experiment. A corollary to Dorn's dictum would be that the analysis of an observational study should reflect the analysis of a controlled experiment in a similar context.

Propensity score methods are a popular tool for the analysis of observational studies. The use of these methods allows one to reduce the effect of the confounding that can occur because of differences in the distribution of measured baseline characteristics between treatment groups. Similar to randomization, propensity score methods remove the effect of confounding by comparing outcomes in treated and untreated subjects who have a similar distribution of measured baseline covariates. Furthermore, these methods allow one to separate the design of an observational study from the analysis of an observational study [8]. Propensity score methods are often applied incorrectly when estimating the effect of treatment on time-to-event outcomes [9, 10]. Common errors include the use of inappropriate statistical tests and the failure to correctly assess whether the specification of the propensity score model had induced acceptable balance in baseline covariates between treated and untreated subjects. Just as the CONSORT statement has improved the reporting of RCTs [11], our objective was to describe methods to improve the conduct and reporting of studies that use propensity score methods to estimate the effect of treatment on time-to-event outcomes using observational data. In particular, we describe how both absolute and relative measures of treatment effect can be estimated.

2. Background

We provide a brief background on the potential outcomes framework, average treatment effects (ATEs), and marginal versus conditional treatment effects. Understanding these concepts allows one to understand differences between conventional regression adjustment and propensity score methods and between different propensity score methods.

2.1. The potential outcomes framework

In a setting with two possible treatments, the potential outcomes framework assumes that each subject has a pair of potential outcomes: $\mathrm{Y}_i(0)$ and $\mathrm{Y}_i(1)$, the outcomes under the control and the active treatment, respectively [12]. However, each subject receives only one of the treatments. Let Z denote the treatment received ($\mathrm{Z} = 0$ for control treatment versus $\mathrm{Z} = 1$ for active treatment). Thus, only one outcome, Y_i, is observed for each subject: the outcome under the treatment received. In studies with survival or time-to-event outcomes, the potential outcomes would be the survival or event times under each of the two treatments.

2.2. Average treatment effects

For each subject, the effect of treatment is defined to be $\mathrm{Y}_i(1)$-$\mathrm{Y}_i(0)$: the difference between the two potential outcomes. The *ATE* is $\mathrm{E}[\mathrm{Y}_i(1)\text{-}\mathrm{Y}_i(0)]$, the average effect of moving an entire population from untreated to treated [13]. A related measure of effect is the average treatment effect for the treated (ATT), $\mathrm{E}[Y(1) - Y(0)|\mathrm{Z} = 1]$, which is the average effect of treatment on those subjects who ultimately received the treatment [13]. Under randomization, the ATE is equal to $\mathrm{E}[Y_i(1)] - \mathrm{E}[Y_i(0)]$ [14]. Therefore, an unadjusted analysis in an RCT allows for unbiased estimation of the average treatment effect at the population level.

In an observational study, there is no reason to expect the ATE and the ATT to coincide. The choice between estimating the ATE or the ATT will often depend on the study question. For instance, when comparing outcomes between two different medications used to treat the same condition, one will often be interested in the ATE, because all patients could easily receive either medication. However, when comparing the effect of a cardiovascular rehabilitation program, the ATT may be of greater interest, because not all eligible patients are likely to elect to participate in the program.

In survival analysis, the effect of treatment for a given subject is the difference between the two potential outcomes. Thus, the average treatment effect denotes the mean difference in survival time because of the treatment. However, biomedical researchers are often more interested in the relative effect of treatment on the hazard of the occurrence of the outcome and in the absolute difference in the probability of the occurrence of the outcome within a specified duration of follow-up time. To reflect this preference, we modify the definitions of the ATE and the ATT for the current study. One can conceptualize two potentially observable survival curves, each of which is a function of the potential outcomes. These would represent survival curves in two identical populations, except that in the first population all subjects were untreated, while in the second population all subjects were treated. Two different metrics could be used for comparing these potentially observable survival curves and thereby quantify the effect of treatment on survival. First, one could estimate the absolute difference in the probability of the occurrence of the outcome between the two potentially observable survival curves. Second, one could pool the two sets of potential outcomes and regress the hazard of the occurrence of the outcome on an indicator variable denoting treatment status. This would allow for estimation of a relative effect of treatment on the hazard of the occurrence of the outcome. We will, in a slight modification of terminology, refer to these measures as measures of the ATE. One could then restrict the above analyses to the set of potential outcomes for those subjects who were ultimately treated. We will refer to the resultant measures as measures of the ATT.

The concept of average treatment effects is important, because some propensity score methods allow one to estimate the ATE, whereas others allow one to estimate the ATT. Understanding which estimate of effect is of primary interest can guide the analyst in selecting the appropriate propensity score method.

2.3. Marginal versus conditional treatment effects and collapsibility

A conditional treatment effect is the average effect, at the individual level, of changing a subject's treatment status from untreated to treated. An estimate of the average effect, at the individual level, is often attained by 'smoothing' the effect across all subjects in the sample. In practice, this is carried out using a regression model in which the outcome is regressed on an indicator variable denoting treatment status and a set of baseline covariates. When such a model has been fit, the regression coefficient for the treatment status indicator variable denotes (possibly after an appropriate transformation) the conditional treatment effect. For instance, after exponentiation, the regression coefficient derived from an adjusted Cox proportional hazards regression model denotes the conditional hazard ratio: the relative change in the hazard of the occurrence of a time-to-event outcome because of treatment. Thus, estimates of treatment effects derived from regression models are conditional effects: the average effect, at the individual level, if a subject's status were changed from untreated to treated.

While conditional effects denote an average effect at the individual level, marginal effects denote an effect at the population level. The marginal treatment effect is the difference in outcomes between two populations that are identical in all respects, except that in one population everyone is treated, while in the second population everyone is untreated. From this definition, one can see that randomized trials are estimating marginal treatment effects.

For a specific population there are multiple conditional effects, one for each of the possible set of covariates that are included in the regression model. In contrast to this, for a specific population, there is only one marginal effect. However, when considering different populations, each could have its own marginal effect. While there are multiple conditional effects, we would argue that the one derived from the true outcomes regression model is of primary interest.

A measure of treatment effect is said to be collapsible if, in the absence of confounding, the conditional and marginal measures of effect coincide [15]. Differences in means and risk differences are collapsible, while odds and hazard ratios are not collapsible [15–17]. The relative risk is not collapsible unless subject-specific relative risks are uniform [15]. Thus, on average, in an RCT, the crude difference in means will coincide with the adjusted difference in means. However, this is not true for odds ratios or hazard ratios [17]. Neuhaus and Jewell [18] demonstrated that for binary outcomes and a logistic

regression model, the adjusted odds ratio will be systematically further from the null than the marginal odds ratio.

Understanding the concepts of marginal and conditional estimates of treatment effect is important, because propensity score methods estimate marginal effects, whereas conventional regression adjustment estimates conditional effects. Knowing which measures of effect are collapsible will provide understanding about the settings in which regression adjustment and propensity score methods would be expected to result in the same estimates of treatment effect.

3. Propensity score methods and survival outcomes

The propensity score is the probability of receiving the active treatment ($Z = 1$ vs. $Z = 0$), conditional on observed baseline covariates ($\mathbf{X}$): $e_i = \Pr(Z_i = 1 \mid \mathbf{X}_i)$ [19]. It is a balancing score: conditional on the propensity score, the distribution of measured baseline covariates is expected to be the same in treated and untreated subjects. Four different propensity score methods are used for reducing the effects of confounding when estimating the effects of treatment on outcomes: propensity score matching, stratification on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score [19–21]. Similar to RCTs, propensity score methods allow one to estimate marginal, rather than conditional measures of treatment effect [22]. The reason for this can be clearly seen for matching, stratification, and weighting: one is comparing average outcomes between samples of treated and untreated subjects who have the same distribution of observed baseline covariates.

There is a lack of consensus in the literature as to which variables one should include in the propensity score model. Rosenbaum [23] suggests that one address the issue of variable selection for the propensity score model by asking 'which covariates do you wish to balance by matching on the propensity score' (p. 356). To do so, we would suggest that one could imagine examining a table comparing baseline covariates of subjects in different treatment arms of an RCT and selecting those baseline covariates about which one would be concerned if baseline imbalance existed. This suggests that one seek to include those variables that are prognostic of the outcome. A simulation study found that this approach to variable selection worked very well in the context of propensity score matching [24]. We suggest that the identification of potentially prognostically important covariates be based on subject-matter expertise and a review of the existing literature, rather than on formal statistical hypothesis testing in the study sample.

3.1. Propensity score matching

Propensity score matching entails forming matched sets of treated and untreated subjects who share a similar value of the propensity score [19, 25]. While one-to-one matching, in which pairs of treated and untreated subjects are formed, appears to be the most common approach to propensity score matching, other approaches are possible [26–28]. Once a matched sample has been formed, the treatment effect can be estimated by directly comparing outcomes between treated and untreated subjects in the matched sample.

We would argue that, in general, variance estimation should account for the matched nature of the propensity score matched sample. Matched subjects have a similar value of the propensity score. Rosenbaum and Rubin demonstrated that subjects with the same propensity score have the same distribution of observed baseline covariates [19]. Thus, matched subjects will, on average, have baseline covariates that are more similar than would two randomly selected subjects. Because baseline covariates are related to the outcome (otherwise there would be no confounding), this implies the existence of a within-matched set correlation in outcomes: matched subjects are likely to display a greater similarity in outcome compared with two randomly selected subjects. We suggest that variance estimation in the propensity score matched sample should account for the lack of independence in outcomes that has been induced by matching. However, we acknowledge that there is not universal agreement on the need to account for matching when estimating significance levels [29]. However, several simulation studies have found that ignoring the paired nature of the matched sample constructed using caliper matching can result in incorrect significance levels, in confidence intervals that do not have correct coverage rates, and in estimates of standard error that overestimate the sampling variability of estimated treatment effect [30–33].

Pair matching permits estimation of the ATT. This can be seen because one has constructed a sample of untreated subjects whose only systematic difference from the sample of treated subjects is the absence of treatment. One can thus compare outcomes between the treated sample and a sample of

untreated subjects that, at baseline, appears to be identical to the treated subjects. By focusing on the effect of treatment in a sample of subjects who resemble the treated subjects, one is estimating the ATT. Pair matching requires that the sample of untreated subjects be larger than the sample of treated subjects. Ideally, there should be substantially more untreated subjects than treated subjects. While pair matching only permits estimation of the ATT, full matching permits estimation of both the ATE and the ATT, depending on how the matched sets are weighted [29].

A common implementation of pair matching is greedy nearest neighbor matching within specified calipers of the propensity score [25]. Using this approach, a treated subject is selected. This treated subject is then matched with the untreated subject whose propensity score is closest to that of the treated subject, subject to the constraint that the differences between their propensity scores is less than a specified maximum (the caliper distance). In practice, one often matches on the logit of the propensity score using a caliper that is defined as a proportion of the standard deviation of the logit of the propensity score [25].

Conventional propensity score matching allows one to estimate the ATT: one is making inferences about the effect of treatment in the population of subjects who received treatment. This is a well-defined population consisting of all subjects who received the treatment. A limitation to caliper matching is the potential for incomplete matching. This occurs when some treated subjects are excluded from the matched sample because there are no available untreated subjects within the specified caliper distance of some of the treated subjects. When incomplete matching has occurred, one is trying to make inferences about the effect of treatment in all patients who received treatment, using a subset of these treated patients. If the unmatched treated subjects differ systematically from the matched treated subjects, it is possible that the estimate of the ATT is biased. Frequently, the unmatched treated subjects will be those subjects with the highest propensity score: those subjects who look like the most likely candidates for treatment. Because matching allows one to estimate the effect of treatment in the treated subjects, it is unclear to what population the estimate applies when some treated subjects have been excluded from the matched sample. Unlike the entire sample of treated subjects, the sample of matched treated subjects may be difficult to describe or define formally.

Alternatives to caliper matching include greedy nearest neighbor matching and optimal matching [25, 34]. The former is similar to caliper matching, except that one removes the constraint that matched subjects must have propensity scores whose differences are less than a specified maximum. Optimal matching forms matched sets such that the total within-pair difference in the propensity score is minimized.

There is likely a tradeoff between different types of bias when choosing between different methods for pair matching. The use of nearest neighbor matching or optimal matching eliminates bias due to incomplete matching, because all treated subjects will be included in the matched sample (assuming that the number of untreated subjects is at least as large as the number of treated subjects). However, their use may result in the matching of more dissimilar subjects, and thus the estimated treatment effect may be contaminated by residual confounding. Caliper matching should result in the elimination of a greater degree of the systematic differences between treated and untreated subjects, but may introduce bias due to incomplete matching.

When outcomes are time-to-event in nature, Kaplan–Meier survival curves can be estimated for treated and untreated subjects in the propensity score matched sample. The estimated survival curves allow one to directly compare survival between treatment groups in the matched sample. On the basis of our arguments previously, we would suggest that it is inappropriate to treat the sample of matched treated subjects and the sample of matched untreated subjects as two independent samples. Thus, we think that, while the log-rank test is frequently used for testing the equality of survival curves in propensity score matched samples [9, 10, 35], such an approach is inappropriate, because it requires that the samples be independent of one another [36, 37]. Instead, the stratified log-rank test can be used to compare the equality of the survival curves in matched samples [36]. Extrapolating from studies that used Monte Carlo simulations to examine the impact of not accounting for matching on other statistical tests, the use of the log-rank test will likely result in type I error rates that are artificially low [31, 32]. However, this requires confirmation in subsequent research.

One may estimate the relative change in the hazard of the outcome by regressing survival on treatment status by using a univariate Cox proportional hazards model. To account for the matched nature of the sample, one can use a robust variance estimator that accounts for the clustering within matched sets [38]. Another approach that has been used is to stratify on the matched sets [39]. However, recent research has demonstrated that, while the former approach allows for unbiased estimation of marginal hazard ratios,

the latter approach results in biased estimation of marginal hazard ratios [30]. The former approach will result in an estimated hazard ratio that is equivalent to that obtained by a conventional Cox proportional hazards model that does not account for clustering and is a true marginal model. The latter approach appears to result in a conditional estimate of effect, because one is conditioning on the matched pairs.

3.2. *Inverse probability of treatment weighting using the propensity score*

Inverse probability of treatment weighting uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. Let Z_i be an indicator variable denoting whether or not the ith subject was treated; furthermore, let e_i denote the propensity score for the ith subject. Weights can be defined as $w_i = \frac{Z_i}{e_i} + \frac{(1-Z_i)}{1-e_i}$. Inverse probability of treatment weighting was first proposed by Rosenbaum [20] as a form of model-based direct standardization. IPTW using the propensity score belongs to a larger class of models called marginal structural models [40] that allow one to account for time-varying confounders when estimating the effect of time-varying exposures.

Cole and Hernán [41] describe a method to estimated adjusted survival curves in the weighed sample, while Xie and Liu [42] describe an adjusted Kaplan–Meier estimate for survival curves in the weighted sample. The latter also proposed a modified log-rank test appropriate for use with weighted samples. Both Cole and Hernán [41] and Joffe *et al*. [43] describe how regression models can be weighted by the inverse probability of treatment to estimate causal effects of treatments. Variance estimation should account for the weighted nature of the synthetic sample, with robust variance estimation commonly being used to account for the sample weights. The relative change in the hazard of the outcome can be estimated using a Cox proportional hazards model in which survival is regressed on an indicator variable denoting treatment status.

Treated subjects with a very low propensity score or untreated subjects with a high propensity score will have large weights. Because of the instability that can be induced by very large weights, Cole and Hernan [41] have recommended that stabilized weights be used. Stabilized weights can be obtained by multiplying the inverse probability of treatment weight by the marginal probability of receiving the actual treatment received. Another alternative to address potential instability due to very large weights is to use trimmed weights, in which weights that exceed a specified threshold are truncated to the threshold value [44].

The weights described previously allow one to estimate the ATE, because the weights are designed to use the overall sample as the reference population. Alternatively, using the following weights allow one to estimate the ATT: $w_i = Z_i + \frac{e_i(1-Z_i)}{1-e_i}$ [45]. With the use of these weights, treated subjects receive a weight of 1, while untreated subjects receive a weight of $\frac{e_i}{1-e_i}$. Thus, the reference population is the sample of treated subjects: both the treated and untreated subjects are weighted so that the distribution of baseline covariates in each of the two samples is the same as in the sample of treated subjects.

3.3. *Other propensity score approaches*

Stratification on the propensity score and covariate adjustment using the propensity score are the two other propensity score methods [19]. With covariate adjustment using the propensity score, one regresses the outcome on the propensity score and an indicator variable denoting treatment selection. With time-to-event outcomes, a Cox proportional hazards model would be used to regress the hazard of the occurrence of the outcome on the propensity score and an indicator variable denoting treatment status. This approach has been shown to result in biased estimation of marginal hazard ratios [30]. Furthermore, it has also been shown to result in a biased estimate of the conditional hazard ratio that would result from adjusting for all the prognostically important covariates in a multivariable Cox regression model [46].

Stratification on the propensity score involves stratifying subjects into mutually exclusive subsets based on their estimated propensity score. In practice, analysts often use five subclasses on the basis of the quintiles of the estimated propensity score. When estimating linear treatment effects, stratum-specific estimates of effect are obtained. These stratum-specific estimates are then pooled to obtain an overall estimate of treatment effect. There are three ways in which one could estimate a hazard ratio using stratification on the propensity score. First, one can estimate stratum-specific Cox regression models in which survival is regressed on treatment selection. The stratum-specific log-hazard ratios are then pooled or averaged to obtain a pooled hazard ratio. Second, one can regress survival on an indicator

variable denoting treatment status and a categorical variable denoting the propensity score strata. Third, one can regress survival on an indicator variable denoting treatment status and stratify on the propensity score strata, thereby allowing the baseline hazard to vary across strata. While stratification performs well for estimating linear treatment effects [19], it results in biased estimation of marginal hazard ratios [30]. Furthermore, one implementation of stratification has also been shown to result in a biased estimate of the conditional hazard ratio that would result from adjusting for all the prognostically important covariates in a multivariable Cox regression model [46]. It appears that each of these approaches results in an estimate of a conditional hazard ratio, rather than a marginal hazard ratio. Further research is required to determine how these conditional hazard ratios differ from that obtained by adjusting for the prognostically important covariates in a conventional Cox regression model.

Because our focus is on methods that allow estimation of both marginal survival curves and marginal hazard ratios, we do not consider these two propensity score methods further in this study.

3.4. Comparison of different propensity score methods

We provide a brief comparison of the four different propensity score methods, without restricting our attention to estimating the effects of treatment on survival outcomes. It has been suggested that matching and stratification may be preferable to IPTW using the propensity score and covariate adjustment using the propensity score because the latter two directly use the estimated propensity score and may thus be more adversely affected by misestimation or instability in the estimated propensity scores [47]. In contrast, the former two approaches use the propensity score for stratifying or matching, but the propensity score is not directly involved in estimating the treatment effect. An additional criticism of covariate adjustment using the propensity score is that it requires the assumption that the outcomes regression model has been correctly specified. While balance diagnostics have been described for covariate adjustment using the propensity score, these diagnostics are less transparent than comparable diagnostics for the other three approaches [48, 49]. Furthermore, matching, stratification, and weighting allow one to separate the design of an observational study from the analysis of an observational study [8]. In a series of Monte Carlo simulations, propensity score matching and IPTW using the propensity score were found to induce better balance on baseline covariates compared with stratification on the propensity score and covariate adjustment using the propensity score [50]. While conventional propensity score matching allows one to estimate the ATE, stratification and IPTW permit estimation of either the ATE or the ATT, depending on how the strata or subjects are weighted. Pair matching on the propensity score requires that the number of untreated subjects be larger (and preferably substantially larger) than the number of treated subjects. Thus, matching will not perform well when the two samples are of approximately equal size or when the number of treated subjects is larger than the number of untreated subjects. The other three propensity score methods do not suffer from this limitation. Finally, the relative performance of these methods for estimating risk differences and marginal hazard ratios has been examined in greater detail elsewhere [51].

3.5. Sensitivity analyses

The critical assumption in propensity score analyses is that of no unmeasured confounding, that one has measured all the variables that influence treatment selection [19]. Analyses have been developed to examine the sensitivity of the observed findings to the presence of unmeasured confounding variables [34, 52]. The methods described in the latter reference assume that there is an unmeasured covariate that increases the odds of treatment assignment. For each specified odds ratio for this unmeasured confounder, one can determine a range of significance levels or P-values for the treatment effect, had the unmeasured confounder been accounted for. The boundaries of this range occur when the unmeasured confounder almost perfectly predicts the outcome. Within matched samples, such sensitivity analyses can be conducted for all sign-score tests [34]. Because the stratified log-rank test is a sign-score test [36], this sensitivity analysis can be applied to the comparison of Kaplan–Meier survival curves in the propensity score matched sample.

4. Case study

4.1. Data source

The sample consisted of patients hospitalized with acute myocardial infarction at 103 acute care hospitals in Ontario, Canada between April 1, 1999 and March 31, 2001. Data on patient history, cardiac risk

factors, comorbid conditions and vascular history, vital signs, and laboratory tests were obtained by retrospective chart review by trained cardiovascular research nurses. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment study [53, 54].

We restricted our sample to patients who survived to hospital discharge. The exposure of interest was whether the patient received a prescription for a statin lipid lowering agent at hospital discharge. We excluded patients with missing data on important baseline clinical covariates. Patient records were linked to the Registered Persons Database using encrypted health card numbers, to allow us to determine the date of death for each patient. Patients were followed for up to 8 years post-discharge. These data provide extended follow-up on a sample of subjects used in a prior article on propensity score methods [55]. Patients who survived to 8 years post-discharge had their survival times treated as censored observations.

The study sample for the current case study consisted of 9107 subjects of whom 3049 (33.5%) received a statin prescription at hospital discharge. We considered 31 baseline characteristics that are grouped in the following categories: demographic characteristics, vital signs on admission, presenting signs and symptoms, classical cardiac risk factors, laboratory tests, cardiac history, and comorbid medical conditions. Baseline characteristics of the study sample are described in Table I. There were systematic differences in baseline characteristics between treated and untreated patients in the overall sample, with treated subjects tending to be younger and healthier than untreated subjects. Ten variables had standardized differences that exceeded 0.10 [49]. Many researchers use a threshold of 0.1 to indicate imbalance in baseline covariates that is of potential concern [56]. The treatment-risk paradox, in which subjects at a decreased risk of death are more likely to receive treatment, has been previously observed and discussed in the context of pharmacological treatment of patients with cardiovascular disease [57, 58].

A propensity score model was estimated using a logistic regression model in which the treatment status was regressed on the 31 baseline variables described previously. The relationship between each continuous variable and the log-odds of statin prescribing was modeled using restricted cubic splines with five knots [59].

4.2. Propensity score matching

Pairs of treated and untreated subjects were matched on the logit of the propensity score using a caliper of width equal to 0.2 of the standard deviation of the logit of the propensity score, as this caliper has been shown to be optimal in a range of settings [60]. Two thousand four hundred and twenty-three (79.5%) of the 3049 treated subjects were matched to an untreated subject with a similar propensity score. Every analysis conducted using propensity score matching should report the method by which matched pairs were formed and the percentage of treated subjects who were included in the matched sample. Approximately 20% of treated subjects were excluded from the matched sample constructed using caliper-matching because no appropriate untreated subject was identified as a match. To address potential biases because of incomplete matching, we constructed a second matched sample using optimal matching.

4.3. Balance diagnostics

An essential component to any propensity score analysis is an assessment of the similarity of baseline covariates between treated and untreated subjects in the matched sample or in the sample weighted by the inverse probability of treatment. We examined balance in baseline covariates using standardized differences. Baseline balance in the 31 covariates is summarized for each method in Table II. Baseline characteristics of treated and untreated subjects in the matched sample constructed using caliper matching are reported in Table I. In examining Tables I and II, one notes that the systematic differences between treated and untreated subjects in the original sample have been substantially reduced or eliminated in the matched sample. In the matched sample constructed using caliper matching, the absolute standardized differences for the 31 baseline covariates ranged from a low of 0 to a high of 0.042. In matched sample obtained using optimal matching, the largest absolute standardized difference was 0.425 (history of hyperlipidemia). The remaining standardized differences were all less than 0.08. In examining Table II, one observes that differences between treatment groups have been reduced by matching or weighting using the propensity score.

The comparisons of means and prevalences and the reporting of standardized differences can be complemented by graphical methods of assessing balance [49]. One could compare the distribution of continuous variables between the two treatment groups using quantile–quantile plots, nonparametric

Table I. Comparison of baseline characteristics between treated and untreated subjects in the original sample and in the propensity score matched sample.

	Original sample			Matched sample (caliper matching)		
Baseline variable	Statin: No (6058)	Statin: Yes (3049)	Standardized difference	Statin: No (2449)	Statin: Yes (2449)	Standardized difference
Demographic characteristics						
Age	68.1 ± 13.8	63.4 ± 12.4	0.355	63.6 ± 12.6	63.8 ± 12.6	0.013
Female	2241 (37.0%)	887 (29.1%)	0.167	738 (30.5%)	736 (30.4%)	0.002
Presenting signs and symptoms						
Acute congestive heart failure/ pulmonary edema	316 (5.2%)	122 (4.0%)	0.057	94 (3.9%)	94 (3.9%)	0.000
Classic cardiac risk factors						
Family history of heart disease	1763 (29.1%)	1177 (38.6%)	0.204	888 (36.6%)	896 (37.0%)	0.007
Diabetes	1562 (25.8%)	774 (25.4%)	0.009	631 (26.0%)	612 (25.3%)	0.018
Hyperlipidemia	1138 (18.8%)	1761 (57.8%)	0.910	1115 (46.0%)	1136 (46.9%)	0.017
Hypertension	2683 (44.3%)	1453 (47.7%)	0.068	1,121 (46.3%)	1,109 (45.8%)	0.010
Current smoker	2004 (33.1%)	1070 (35.1%)	0.043	894 (36.9%)	875 (36.1%)	0.016
Cardiac history and comorbid conditions						
CVA/TIA	610 (10.1%)	237 (7.8%)	0.079	181 (7.5%)	196 (8.1%)	0.023
Angina	1871 (30.9%)	1086 (35.6%)	0.101	833 (34.4%)	822 (33.9%)	0.010
Cancer	191 (3.2%)	73 (2.4%)	0.045	64 (2.6%)	60 (2.5%)	0.010
Dementia	243 (4.0%)	33 (1.1%)	0.171	26 (1.1%)	29 (1.2%)	0.012
Previous AMI	1254 (20.7%)	799 (26.2%)	0.132	566 (23.4%)	562 (23.2%)	0.004
Asthma	338 (5.6%)	166 (5.4%)	0.006	110 (4.5%)	132 (5.4%)	0.042
Depression	441 (7.3%)	192 (6.3%)	0.039	154 (6.4%)	159 (6.6%)	0.008
Hyperthyroidism	71 (1.2%)	40 (1.3%)	0.013	37 (1.5%)	28 (1.2%)	0.032
Peptic ulcer disease	345 (5.7%)	156 (5.1%)	0.025	127 (5.2%)	120 (5.0%)	0.013
Peripheral vascular disease	430 (7.1%)	220 (7.2%)	0.005	187 (7.7%)	178 (7.3%)	0.014
Previous coronary revascularization	433 (7.1%)	411 (13.5%)	0.220	267 (11.0%)	264 (10.9%)	0.004
Congestive heart failure (chronic)	275 (4.5%)	91 (3.0%)	0.079	86 (3.5%)	71 (2.9%)	0.035
Stenosis	96 (1.6%)	35 (1.1%)	0.037	21 (0.9%)	30 (1.2%)	0.036
Vital signs on admission						
Systolic blood pressure	148.7 ± 31.6	149.3 ± 30.1	0.021	149.8 ± 30.8	149.0 ± 30.4	0.025
Diastolic blood pressure	83.6 ± 18.6	84.5 ± 18.0	0.047	84.4 ± 18.6	84.5 ± 18.2	0.010
Heart rate	84.6 ± 24.3	81.7 ± 23.0	0.121	81.3 ± 22.5	81.7 ± 22.5	0.020
Respiratory rate	21.2 ± 5.7	20.3 ± 4.8	0.166	20.4 ± 4.9	20.3 ± 4.9	0.005
Results of initial laboratory tests						
White blood count	10.3 ± 4.9	10.0 ± 4.4	0.065	10.1 ± 4.4	10.1 ± 4.6	0.008
Hemoglobin	137.5 ± 19.3	140.6 ± 16.9	0.167	140.6 ± 17.4	140.5 ± 17.4	0.003
Sodium	138.9 ± 3.9	139.2 ± 3.3	0.079	139.2 ± 3.7	139.2 ± 3.3	0.011
Glucose	9.4 ± 5.1	9.2 ± 5.3	0.037	9.3 ± 5.6	9.2 ± 5.2	0.013
Potassium	4.1 ± 0.6	4.1 ± 0.5	0.061	4.1 ± 0.5	4.1 ± 0.5	0.023
Creatinine	105.7 ± 65.4	99.9 ± 50.0	0.096	100.4 ± 54.3	101.1 ± 54.2	0.014

CVA, cerebral vascular accident; TIA, transient ischemic attack; AMI, acute myocardial infarction. Continuous variables are reported as means ± standard deviations. Dichotomous variables are reported *N* (%). The propensity score matched sample was constructed greedy nearest neighbor matching on the logit of the propensity score using calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score.

Statistics in Medicine

Table II. Standardized differences comparing baseline covariates between treated and untreated subjects using different propensity score methods.

	31 baseline covariates					55 pairwise interactions of continuous covariates				
Propensity score method	Minimum	25th percentile	Median	75th percentile	Maximum	Minimum	25th percentile	Median	75th percentile	Maximum
Full sample (unmatched and unweighted)	0.005	0.038	0.068	0.169	0.875	0.003	0.046	0.092	0.155	0.348
Caliper matching	0	0.008	0.013	0.020	0.042	0.001	0.005	0.011	0.016	0.033
Optimal matching	0.001	0.013	0.026	0.046	0.425	0.003	0.012	0.026	0.047	0.061
Weighting using the inverse probability of treatment weights	0.002	0.007	0.014	0.018	0.039	0.001	0.004	0.010	0.016	0.036

density plots, or empirical cumulative distribution functions. Because of space constraints, we do not report these additional methods here. The objective of balance diagnostics is to examine whether adequate balance on baseline covariates has been induced by the current specification of the propensity score model. Adequate balance has been achieved when potentially prognostically important covariates have been balanced between treated and untreated subjects in the matched or weighted sample. As described by Rosenbaum and Rubin [61], one may need to iteratively modify the specification of the propensity score model in order to induce acceptable balance on baseline covariates. It has been suggested that one attempt to achieve balance similar to what would be expected in a similarly sized RCT [49].

When using propensity score matching, several authors have discouraged analysts from using statistical significance testing to compare baseline characteristics with the matched sample [10, 62]. However, we should note that not all authors discourage this practice. Hansen and Bowers [63] argue that significance testing has a role in balance assessment and have developed omnibus tests of balance for randomized studies that can also be used with matching or stratification on the propensity score.

4.4. Estimating survival curves and absolute differences in survival using propensity score methods

Crude Kaplan–Meier survival curves for treated and untreated subjects in the full original sample are reported in the top-left panel of Figure 1(the two survival curves were significantly different from one another; log-rank test: $P < 0.0001$). Kaplan–Meier survival curves for treated and untreated subjects in the two propensity score matched samples are described in the top-center and top-right panels of Figure 1 (stratified log-rank test: $P = 0.0104$ and 0.0117, respectively). Survival curves in the sample weighted using the ATE weights and the sample weighted using the ATT weights are reported in the bottom-left and bottom-center panels, respectively (adjusted log-rank test: $P < 0.0001$).

Using the estimated survival curves, one can estimate the absolute reduction in the probability of death within 8 years of discharge because of statin prescribing. The absolute reduction in the probability of death within 8 years due to statin prescribing was 0.039, 0.023, 0.055, and 0.018 for the matched (caliper matching), matched (optimal matching), weighted (ATE weights), and weighted (ATT weights), respectively. The numbers needed to treat were 26, 45, 18, and 57, respectively. While these absolute risk reductions were for 8 years of follow-up, they could easily be computed for any duration of length less than 8 years.

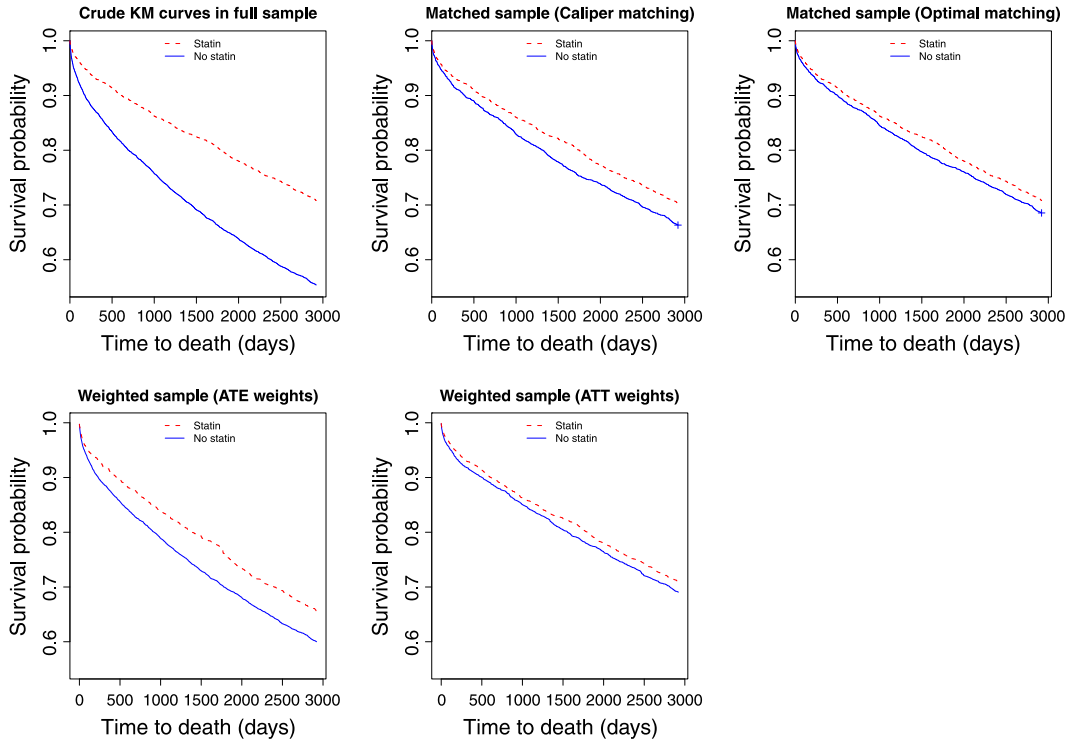

Figure 1. Kaplan–Meier survival curves obtained using different propensity score methods.

4.5. Estimating relative survival effects using propensity score methods

In the matched sample constructed using caliper matching, a Cox model was used to regress survival on treatment status, with a robust variance estimator used to account for the clustering within matched sets. The estimated hazard ratio was 0.855 (95% CI: 0.779–0.939). Thus, receipt of a statin prescription reduced the hazard of post-discharge death by 14.5%. In the matched sample obtained using optimal matching, the estimated hazard ratio was 0.909 (95% CI: 0.834–0.991). A Cox proportional hazards regression model was used to regress survival on treatment status in the sample weighted using the ATE weights. A robust, sandwich-type variance estimator was used to account for the weighted nature of the sample. The estimated hazards ratio was 0.815 (95% CI: 0.745–0.890). As a sensitivity analysis, we repeated the previous analysis using stabilized ATE weights. The estimated hazards ratio was 0.815 (95% CI: 0.746–0.890). When the sample was weighted using the ATT weights, the estimated hazard ratio was 0.927 (95% CI: 0.846–1.016). Thus, among treated subjects, statin prescribing reduced the hazard of death by 7.3%.

4.6. Sensitivity to unmeasured confounders

We conducted an analysis to examine the sensitivity of our findings to the assumption of no unmeasured confounders. We assumed that there was an unmeasured confounding variable that increased the odds of exposure by Γ. In the sample obtained using caliper matching, we excluded those matched pairs in which the subject with the shortest observation time was a censored observation. This resulted in the retention of 1230 matched pairs. Among these pairs, there were 660 pairs in which the untreated subject died first and there were 570 pairs in which the treated subject died first. Let p^+ denote $\Gamma/(1+\Gamma)$ and p^- denote $1/(1+\Gamma)$. Then, the true significance level, if one were to account for the unmeasured confounding variable, would lie in the interval $\left(\sum_{a=660}^{1230}\binom{1230}{a}(p^+)^a(1-p^+)^{1230-a}, \sum_{a=660}^{1230}\binom{1230}{a}(p^-)^a(1-p^-)^{1230-a}\right)$. If $\Gamma = 1.05$, then the interval is (0.000344, 0.046147). Thus, if an unmeasured confounder increased the odds of exposure by 5%, the effect of treatment on survival would remain statistically significant if one were able to account for this confounder. However, if $\Gamma = 1.10$, then the interval is (0.00001, 0.19257). Thus, if an unmeasured confounder increased the odds of exposure by 10% and was almost perfectly associated with death, accounting for this unmeasured confounder could render statistically nonsignificant the effect of treatment on survival. Thus, our study may be relatively susceptible to unmeasured confounding. We recall that the P-value for the stratified log-rank test in the caliper-matched sample was 0.0104. Thus, a small P-value in the primary analysis cannot be taken as an indication that the study is insensitive to unmeasured confounders.

4.7. Conditional effects

Throughout this paper, we have focused on the estimation of marginal treatment effects. The rationale for this focus was based on our desire to mimic the types of analyses that are conducted in RCTs with survival outcomes. We have noted that stratification on the propensity score and covariate adjustment using the propensity result in biased estimation of marginal hazard ratios, because they appear to be estimating conditional measures of effect. For comparative purposes, we estimated several different conditional hazard ratios so that these could be compared with the estimates of the marginal hazard ratios obtained previously.

First, we used a Cox proportional hazards model to regress survival on an indicator variable denoting treatment status and the baseline covariates listed in Table 1. The relationship between continuous covariates and the log-hazard of death was represented using restricted cubic splines with five knots. The estimated hazard ratio was 0.805 (95% CI: 0.739–0.977). Second, a univariate Cox proportional hazards model that stratified on the matched pairs was fit to each of the two matched samples. The estimated hazard ratio was 0.864 (95% CI: 0.772–0.966) in the matched sample constructed using caliper matching, while it was 0.878 (95% CI: 0.794–0.972) in the matched sample constructed using optimal matching. Third, Cox regression was used to regress survival on an indicator variable denoting treatment status and the propensity score. The estimated hazard ratio was 0.856 (95% CI: 0.790–0.928). Fourth, two different approaches to stratification on the quintiles of the propensity score were employed. In the first approach, Cox regression was used to regress survival on an indicator variable denoting treatment status and a categorical variable denoting the five propensity-score strata. In the second approach, a univariate Cox

regression model was used to regress survival on an indicator variable denoting treatment status. This model stratified on the quintiles of the propensity score, allowing the baseline hazard function to vary across quintiles. In both cases, the estimated hazard ratio was 0.870 (95% CI: 0.802–0.945).

5. Discussion

The CONSORT statement recommends that, for RCTs with dichotomous outcomes, both relative and absolute measures of treatment effect be reported [1]. In keeping with the intent of the CONSORT statement, a thorough analysis of the effect of treatment on survival time in an RCT includes estimation of both relative and absolute measures of treatment effect. We suggest that the estimation of treatment effects in observational studies should reflect the analyses that would be conducted in a similarly designed RCT. Propensity score methods permit the estimation of survival curves in treated and untreated subjects. These marginal survival curves reflect the survival functions in the population if all subjects were treated or if all subjects were untreated. When propensity score matching or IPTW are used, univariate regression using a Cox proportional hazards model in which survival is regressed on an indicator variable denoting treatment status allows one to estimate the relative reduction in the hazard of an event occurring. Thus, an analysis using these two propensity score methods allows one to report absolute and relative reductions in the likelihood of the occurrence of an event, measures of effect similar to those that would be reported in an RCT. We summarize our recommendations for the use of propensity score methods with time-to-event outcomes in Table III.

We have discouraged the use of propensity score methods that appear to result in conditional estimates of treatment effect. This should not be taken as a suggestion that conditional measures of effect are of less interest than marginal measures of effect. Instead, there are three motivations for this emphasis. First, propensity score methods are intended to provide estimates of marginal treatment effect [22]. Second, both stratification on the propensity score and covariate adjustment using the propensity score can result in biased estimation of the conditional hazard ratio that would be obtained by adjusting for all prognostically important covariates [30]. Third, we have focused on methods that allow for the estimation of treatment effects that are reported in RCTs. In RCTs, regression adjustment of the outcome on treatment status alone allows for the estimation of marginal hazard ratios. As a consequence, we have discouraged the use of stratification on the propensity score and covariate adjustment using the propensity score to estimate hazard ratios. The amplification of the magnitude of the estimated treatment effect when a conditional estimate is reported compared to the marginal effect can have important impacts on

Table III. Recommendations for the use of propensity score methods with time-to-event outcomes.

Objectives	Propensity score matching (pair matching)	Inverse probability of treatment weighting using the propensity score
Estimand	ATT	ATE or ATT depending on the weights selected
Balance assessment	Compare distribution of baseline covariates between treated and untreated subjects in the matched sample.	Compare distribution of baseline covariates between treated and untreated subjects in the sample weighted by the inverse probability of treatment.
Estimate and report survival curves	Estimate Kaplan–Meier survival curves in treated and untreated subjects in the matched sample. Use stratified log-rank test to compare survival curves (stratify on matched sets).	Estimate adjusted Kaplan–Meier survival curves in the weighted sample. Use adjusted log-rank test to compare survival curves.
Estimate and report absolute reduction in the probability of an event occurring	From the estimated survival curves, estimate the absolute difference in the probability of an event occurring within a specified duration of follow-up.	From the estimated marginal survival curves, estimate the absolute difference in probability of an event occurring within a specified duration of follow-up.
Estimate relative change in the hazard of an event occurring	Use Cox proportional hazards model in the matched sample. Regress survival on an indicator variable for treatment selection. Use a robust variance estimator.	Use Cox proportional hazards model in the weighted sample. Regress survival on an indicator variable for treatment selection. Use a robust variance estimator.

ATT, average treatment effect for the treated; ATE, average treatment effect.

medical decision making and policy decisions. For instance, when interpreting the impact of treatment or exposure at the population level, use of conditional estimates can result in overestimates of the benefits of treatments or of the harms of exposures. This may result in subsequent misuse of resources. As previously noted, there are multiple conditional effects for a given population. The conditional hazard ratio obtained by adjusting for the propensity score may differ from the conditional hazard ratio obtained by directly adjusting for the covariates that are associated with the outcome (this was observed to occur in some of our secondary analyses examining estimation of conditional hazard ratios) [46]. It is unclear how the conditional hazard ratios obtained using covariate adjustment using the propensity score or stratification on the propensity score are related to the primary conditional hazard ratio of interest. Further research is required to address this issue. We suspect that in many settings, the primary conditional hazard ratio of interest may be best approximated by regressing the outcome on an indicator variable denoting treatment status and on all measured covariates (assuming a set of a moderate number of covariates) and by using flexible smoothing methods to relate continuous covariates to the log-hazard of the outcome.

When estimating the effect of treatment on survival outcomes, we recommend that researchers use either propensity score matching or inverse probability of treatment weighting using the propensity score. There are advantages and limitations of each of these two methods. An advantage to matching is that it may be perceived to be more transparent than weighting, which relies on the creation of a synthetic weighted sample. Furthermore, weighting may be more sensitive to misspecification of the propensity score [47]. A disadvantage to conventional matching is that one is restricted to estimating the ATT, whereas weighting allows estimation of either the ATE or the ATT, depending on the weights selected. Depending on the context of the study, one of these may be more useful and informative than the other. In some settings, researchers may want to report both measures of effect. Recent research has shown that weighting and matching eliminate systematic differences between treated and untreated subjects to an approximately equivalent degree [50]. An advantage to propensity score weighting is that this method is a subclass of a more general family of models, Marginal Structural Models, that allow one to account for time-varying exposures and time-varying confounders. Thus, weighting generalizes to an approach that allows one to examine more complex study designs and research questions. However, it should be noted that Lu has extended conventional propensity score methods to settings in which treatment is time varying and is not necessarily fixed at baseline [64]. A potential limitation of matching is that, ideally, it requires a pool of potential controls that is at least as large as the set of treated subjects. In some research contexts, the number of treated subjects may exceed the number of untreated subjects. To conclude our comparison of matching and weighting, we would argue that, in general, neither approach is clearly superior to the other, and that the relative strengths and limitations of each approach need to be considered when selecting a method. In some settings, one approach may have clear advantages over the other.

As noted previously, there are tradeoffs in choosing between caliper matching and optimal matching. Caliper matching should result in the elimination of a greater degree of the systematic differences between treated and untreated subjects, but may introduce bias due to incomplete matching. In our primary matched analysis (in which approximately 20% of treated subjects were excluded) the estimated hazard ratio was 0.855, whereas it was 0.909 in the analysis that used optimal matching in which all treated subjects were included. It is instructive to compare the panels for these analyses in Figure 1. The survival curves estimated in the sample constructed using optimal matching are very similar to the survival curves obtained using weighting using the ATT weights. However, the survival curve for untreated subjects in the matched sample constructed using caliper matching is different from the two corresponding curves obtained using the other two methods. In particular, survival was modestly worse for untreated subjects in the caliper-matched sample compared with the untreated subjects in the optimally matched sample and in the sample weighted using the ATT weights. It appears that the use of caliper matching resulted in the comparison of the survival of treated subjects with that of control or untreated subjects with worse prognosis compared with when the other two methods for estimating the ATT were used. These analyses suggest that the results obtained using caliper matching may be subject to more bias compared with the results from the other two ATT analyses.

The CONSORT statement has improved the reporting of RCTs [11]. Adherence to the methods described in the current paper will improve the conduct and reporting of studies that use propensity score methods to estimate the effect of treatment on time-to-event outcomes. When outcomes are time-to-event in nature, we encourage the use of propensity score matching or inverse probability of treatment weighting using the propensity score for the following reasons. First, they permit one to estimate marginal

survival functions for treated and untreated populations. From these survival functions, one can compute the absolute reduction in the probability of an event occurring within a specified duration of follow-up. Second, each of these methods permits estimation of marginal hazard ratios, which allow one to quantify the relative reduction in the hazard of an event occurring in a treated population compared with an untreated population. Use of these methods will allow one to mimic the reporting of RCTs, by permitting the reporting of both absolute and relative measures of effects in observational studies.

Acknowledgements

This study was supported by the Institute for Clinical Evaluative Sciences, which is funded by an annual grant from the Ontario Ministry of Health and Long-term Care. The opinions, results and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by Institute for Clinical Evaluative Sciences or the Ontario Ministry of Health and Long-term Care is intended or should be inferred. This research was supported by operating grant from the Canadian Institutes of Health Research (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation. The Enhanced Feedback for Effective Cardiac Treatment data used in the study was funded by a Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research. These data sets were held securely in a linked, de-identified form and analyzed at the Institute for Clinical Evaluative Sciences.

References

1. Schulz KF, Altman DG, Moher D. Consort 2010 Statement: updated guidelines for reporting parallel group randomised trials. *British Medical Journal* 2010; **340**:c332. DOI: 10.1136/bmj.c332.
2. Austin PC, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. *International Journal of Biostatistics* 2011; **7**(1):Article 6. DOI: 10.2202/1557-4679.1285.
3. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *New England Journal of Medicine* 1988; **318**:1728–1733.
4. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal* 1995; **310**(6977):452–454.
5. Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *Canadian Medical Association Journal* 1995; **152**(3):351–357.
6. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology* 2010; **63**(2):142–153.
7. Rubin DB. *Matched sampling for causal effects*. Cambridge University Press: New York, NY, 2006.
8. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* 2007; **26**(1):20–36.
9. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *Journal of Thoracic and Cardiovascular Surgery* 2007; **134**(5):1128–1135.
10. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine* 2008; **27**(12):2037–2049.
11. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *Journal of the American Medical Association* 2001; **285**:1992–1995.
12. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**:688–701.
13. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004; **86**:4–29.
14. 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**(19):2937–2960.
15. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**(5):761–768.
16. Miettinen OS, Cook EF. Confounding essence and detection. *American Journal of Epidemiology* 1981; **4**:593–603.
17. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; **7**:431–444.
18. Neuhaus JM, Kalbfleish JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; **59**(1):25–35.
19. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
20. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**:387–394.
21. Austin PC. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; **46**:399–424. DOI: 10.1080/00273171.2011.568786.
22. Rosenbaum PR. Propensity score. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). John Wiley & Sons: Boston, 2005; 4267-4272.

23. Rosenbaum PR. *Design of Observational Studies*. Springer-Verlag: New York, NY, 2010.
24. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in Medicine* 2007; **26**(4):734–753.
25. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 1985; **39**:33–38.
26. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 1993; **2**:405–420.
27. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 2004; **99**(467):609–618.
28. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* 2000; **56**(1):118–124.
29. Stuart EA. Matching methods for causal inference: a review and a look forward. *Statistical Science* 2010; **25**(1):1–21.
30. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stastisics in Medicine* 2013; **32**(16):2837–2849. DOI: 10.1002/sim.5705.
31. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *International Journal of Biostatistics* 2009; **5**(1). DOI: 10.2202/1557-4679.1146.
32. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Statistics in Medicine* 2011; **30**(11):1292–1301.
33. Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharmaceutical Statistics* 2012; **11**(3):222–229. DOI: 10.1002/pst.537.
34. Rosenbaum PR. *Observational studies*. Springer-Verlag: New York, NY, 2002.
35. Austin PC. A report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review and suggestions for improvement. *Circulation: Cardiovascular Quality and Outcomes* 2008; **1**:62–67.
36. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag: New York, NY, 1997.
37. Harrington D. Linear rank tests in survival analysis. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). John Wiley & Sons: New York, NY, 2005; 2802–2812.
38. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *Journal of the American Statistical Association* 1989; **84**:1074–1078.
39. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag: New York, 2000.
40. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**(5):550–560.
41. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 2004; **75**:45–49.
42. Xie J, Liu C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 2005; **24**(20):3089–3110.
43. 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.
44. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS One* 2011; **6**(3):e18174.
45. Morgan SL, Todd JL. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* 2008; **38**:231–281.
46. Austin PC, Grootendorst P, Normand SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Statistics in Medicine* 2007; **26**(4): 754–768.
47. Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* 2004; **13**(12):855–857.
48. Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiology and Drug Safety* 2008; **17**(12):1202–1217.
49. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 2009; **28**(25):3083–3107.
50. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making* 2009; **29**(6):661–677.
51. Austin PC. The performance of different propensity score methods for estimating difference in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010; **29**:2137–2148. DOI: 10.1002/sim.3854.
52. Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society - Series B* 1983; **45**:212–218.
53. Tu J, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM. *Quality of Cardiac Care*. Institute for Clinical Evaluative Sciences: Toronto, ON, 2004.
54. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association* 2009; **302**(21):2330–2337.
55. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**(12):2084–2106.
56. Normand ST, Landrum MB, Guadagnoli E, Ayanian JZ, Ryan TJ, Cleary PD, McNeil BJ. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of Clinical Epidemiology* 2001; **54**(4):387–398.

57. Ko DT, Mamdani M, Alter DA. Lipid-lowering therapy with statins in high-risk elderly patients: the treatment-risk paradox. *Journal of the American Medical Association* 2004; **291**:1864–1870.
58. Lee DS, Tu JV, Juurlink DN, Alter DA, Ko DT, Austin PC, Chong A, Stukel TA, Levy D, Laupacis A. Risk-treatment mismatch in the pharmacotherapy of heart failure. *Journal of the American Medical Association* 2005; **294**(10):1240–1247.
59. Harrell, F E Jr. *Regression modeling strategies*. Springer-Verlag: New York, NY, 2001.
60. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics* 2010; **10**:150–161.
61. 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.
62. Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society - Series A (Statistics in Society)* 2008; **171**:481–502.
63. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science* 2008; **23**:219–236.
64. Lu B. Propensity score matching with time-dependent covariates. *Biometrics* 2005; **61**(3):721–728.

