Research Article

Statistics in Medicine

Received 28 November 2011, Accepted 19 November 2012 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5705

The performance of different propensity score methods for estimating marginal hazard ratios

Peter C. Austin[a,b,c*†]

Propensity score methods are increasingly being used to reduce or minimize the effects of confounding when estimating the effects of treatments, exposures, or interventions when using observational or non-randomized data. Under the assumption of no unmeasured confounders, previous research has shown that propensity score methods allow for unbiased estimation of linear treatment effects (e.g., differences in means or proportions). However, in biomedical research, time-to-event outcomes occur frequently. There is a paucity of research into the performance of different propensity score methods for estimating the effect of treatment on time-to-event outcomes. Furthermore, propensity score methods allow for the estimation of marginal or population-average treatment effects. We conducted an extensive series of Monte Carlo simulations to examine the performance of propensity score matching (1:1 greedy nearest-neighbor matching within propensity score calipers), stratification on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score to estimate marginal hazard ratios. We found that both propensity score matching and IPTW using the propensity score allow for the estimation of marginal hazard ratios with minimal bias. Of these two approaches, IPTW using the propensity score resulted in estimates with lower mean squared error when estimating the effect of treatment in the treated. Stratification on the propensity score and covariate adjustment using the propensity score result in biased estimation of both marginal and conditional hazard ratios. Applied researchers are encouraged to use propensity score matching and IPTW using the propensity score when estimating the relative effect of treatment on time-to-event outcomes.

Keywords: propensity score; survival analysis; inverse probability of treatment weighting (IPTW); Monte Carlo simulations; observational study; time-to-event outcomes

1. Introduction

Observational studies are increasingly being used to estimate the effects of treatments, interventions, and exposures on outcomes. These studies allow for the examination of treatment effects in settings in which randomized controlled trials (RCTs) may be unethical or impractical. Furthermore, observational studies allow for the study of treatment efficacy outside of the tightly controlled environment of an RCT, allow for the inclusion of subjects who may have been excluded from RCTs, and allow for the study of rare outcomes and adverse events for which RCTs may have inadequate statistical power. The advantage of RCTs is that random allocation of treatment assignment allows one to obtain an unbiased estimate of the average treatment effect [1]. This is because there will, on average, be no systematic differences in baseline covariates between treatment groups. In contrast, in observational studies, treatment allocation is frequently influenced by subject characteristics. Therefore, there often exist systematic differences between treatment groups in observational studies. We must use statistical methods to remove or minimize the effect of this confounding so that valid inferences on treatment effect can be drawn from observational studies.

[a] *Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*

[b] *Institute of Health Management, Policy and Evaluation, University of Toronto, Toronto, Ontario, Canada*

[c] *Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, 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*

Propensity score methods are increasingly being used to reduce or minimize the confounding that occurs frequently in observational studies of the effect of treatment on outcomes. The propensity score is the probability of treatment assignment conditional on measured baseline covariates [2]. There are four ways of using the propensity score to reduce confounding: matching on the propensity score, stratification on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, and covariate adjustment using the propensity score. These methods are often used in the biomedical literature [3, 4].

Conditioning on the propensity score results in unbiased estimates of linear treatment effects [2]. Thus, when outcomes are continuous, conditioning on the propensity score allows for unbiased estimation of differences in means. However, in biomedical research, outcomes are commonly binary or time to event in nature, rather than continuous [5]. When outcomes are binary, we can estimate the effect of treatment using the risk difference (difference in proportions or absolute risk reduction) (along with the associated number needed to treat, which is the reciprocal of the absolute risk reduction), the relative risk, and the odds ratio. Several studies have examined the performance of different propensity score methods for estimating treatment effects when outcomes are binary [6–9]. Although propensity score methods have frequently been used in the analysis of time-to-event outcomes, there is a paucity of research examining the relative performance of different propensity score methods for estimating hazard ratios.

A conditional effect is the average effect, at the subject level, of moving a subject from untreated to treated. The regression coefficient for a treatment assignment indicator variable from a multivariable regression model is an estimate of a conditional or adjusted effect. In contrast, a marginal effect is the average effect, at the population level, of moving an entire population from untreated to treated [10]. Linear treatment effects (differences in means and differences in proportions) are collapsible: the conditional and marginal treatment effects will coincide. However, when outcomes are binary or time to event in nature, the odds ratio and the hazard ratio are not collapsible [11]. Rosenbaum has noted that propensity score methods allow one to estimate marginal, rather than conditional, effects [12]. There is a paucity of research into the performance of different propensity score methods to estimate marginal treatment effects.

The objective of the current study is to examine the ability of different propensity score methods to estimate marginal and conditional hazard ratios when outcomes are time to event in nature. The paper is structured as follows. In Section 2, we describe different propensity score methods and how they can be used to estimate hazard ratios for survival outcomes. In Section 3, we describe the design and results of an extensive series of Monte Carlo simulations to compare the performance of different propensity score methods to estimate hazard ratios. In Section 4, we summarize our findings and place them in the context of the existing literature.

2. Propensity score methods and survival outcomes

We use the following notation throughout this section. Let Z be an indicator variable denoting treatment status ($Z = 1$ for active treatment of interest and $Z = 0$ for the control treatment), whereas e denotes the estimated propensity score.

2.1. Matching on the propensity score

Matching on the propensity score entails forming matched sets of treated and untreated subjects who have a similar value of the propensity score [13]. The most common implementation of propensity score matching is pair matching or 1:1 matching in which matched pairs of treated and untreated subjects are formed. In the current study, we used greedy nearest-neighbor matching within specified caliper widths to form pairs of treated and untreated subjects matched on the logit of the propensity score [13]. We used calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score as this caliper width has been found to perform well in a wide variety of settings [14].

Once a propensity-score-matched sample had been formed, we estimated the effect of treatment on survival using three different methods. First, we used a Cox proportional hazards regression model to regress survival on an indicator variable denoting treatment status. We used model-based standard errors to estimate 95% confidence intervals. Second, we fitted the same model as before; however, we obtained a robust sandwich estimate of the variance of the regression coefficient that accounted for the clustering within matched sets [15]. Finally, we fitted a univariate Cox model as before; however, this model stratified on matched sets, thereby allowing the baseline hazard function to vary across matched sets.

We carried this out to account for the potential homogeneity of outcomes within matched sets. Cummings, McKnight, and Greenland proposed the use of stratification on matched sets to account for matched cohort designs with time-to-event outcomes [16]. We will refer to these three models as the naïve Cox model, the robust Cox model, and the stratified Cox model, respectively.

2.2. *Stratification on the propensity score*

Stratification (or subclassification) on the propensity score stratifies the entire sample into mutually exclusive subclasses on the basis of the propensity score. A common approach is to define the subclasses using specified quantiles of the propensity score. Using the quintiles of the estimated propensity score to divide the sample into five, approximately equally sized, groups has been shown to eliminate approximately 90% of the bias due to measured confounding variables when estimating a linear treatment effect [2, 17, 18]. We used stratification on the quintiles of the propensity score in the current study given its popularity in the applied literature.

When estimating a linear treatment effect (e.g., a difference in means or difference in proportions), one can estimate stratum-specific treatment effects and then pool or average these stratum-specific effects across the strata [2, 17]. We examined three modifications to this approach for estimating hazard ratios when outcomes are time to event in nature. Each method was based on a Cox regression model with survival as the outcome variable. In the first method, we included two explanatory variables in the Cox model: an indicator variable denoting treatment status and a categorical variable denoting propensity score strata (as a five-level categorical variable). In the second method, we fit five stratum-specific univariate Cox regression models in which survival was regressed on an indicator variable denoting treatment status. We then pooled or averaged the five log-hazard ratios to estimate an overall treatment effect. Third, we fit a univariate Cox model in which we regressed survival on an indicator variable denoting treatment status. The model stratified on the five propensity score strata, thereby allowing the baseline hazard ratio to vary across the propensity score strata. We refer to these three methods as stratification (adjusted), stratification (pooled), and stratification (stratified), respectively.

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

The IPTWs are defined as $(Z/e) + [(1-Z)/(1-e)]$ [19]. Weighting the sample using these weights results in a weighted synthetic sample in which observed baseline covariates are not confounded with treatment assignment. Using these weights allows one to estimate the average treatment effect (ATE). Using weights equal to $Z + [e(1-Z)/(1-e)]$ allows one to estimate the average treatment effect in the treated (ATT) [20]. In the weighted sample (weighted using either the ATE weights or the ATT weights), we used a Cox regression model to regress survival on an indicator variable denoting treatment status and used a robust variance estimator [15, 21].

2.4. *Covariate adjustment using the propensity score*

Rosenbaum and Rubin proposed covariate adjustment using the propensity score in the context of estimating linear treatment effects for continuous outcomes [2]. Using this approach, we regress the outcome on two covariates: an indicator variable denoting treatment status and the propensity score. The regression coefficient associated with the treatment selection indicator represents the effect of treatment. In the current study, we used Cox regression to regress survival time on these two variables. The regression coefficient for the treatment status indicator is the estimated log-hazard ratio.

3. Monte Carlo simulations

We used a series of Monte Carlo simulations to examine the relative performance of different propensity score methods to estimate hazard ratios. Our primary focus was on estimating marginal or population-average hazard ratios. However, as a secondary objective, we also examined estimation of conditional hazard ratios.

3.1. *Monte Carlo simulations—methods*

We simulated data for a setting in which there were 10 baseline covariates (X_1–X_{10}). We simulated these covariates from independent standard normal distributions. Of these 10 covariates, seven affected treatment selection (X_1–X_7), whereas seven affected the outcome (X_4–X_{10}). Furthermore, we allowed covariates to have a weak, moderate, strong, or very strong effect on treatment selection or outcome. For each subject, we determined the probability of treatment selection from the following logistic model: $\text{logit}(p_i) = \alpha_{0,\text{treat}} + \alpha_W x_1 + \alpha_M x_2 + \alpha_S x_3 + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS} x_7$. We selected the intercept of the treatment selection model ($\alpha_{0,\text{treat}}$) so that the proportion of subjects in the simulated sample that were treated was fixed at the desired proportion. The regression coefficients α_W, α_M, α_S, and α_{VS} were set to log(1.25), log(1.5), log(1.75), and log(2), respectively. These were intended to denote weak, moderate, strong, and very strong treatment assignment affects. For each subject, we generated treatment status from a Bernoulli distribution with subject-specific parameter p_i.

We then generated a time-to-event outcome for each subject using a data-generating process for time-to-event outcomes described by Bender *et al.* [22]. For each subject, we defined the linear predictor as $\text{LP} = \beta_{\text{treat}} Z + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS} x_7 + \alpha_W x_8 + \alpha_M x_9 + \alpha_S x_{10}$. For each subject, we generated a random number from a standard uniform distribution: $u \sim \text{U}(0, 1)$. We generated a survival or event time for each subjects as follows: $-\log(u)/\left(\lambda e^{\text{LP}}\right)^{1/\eta}$. We set λ and η to be equal to 0.00002 and 2, respectively. The use of this data-generating process results in a conditional treatment effect, with a conditional hazard ratio of $\exp(\beta_{\text{treat}})$. However, we wanted to generate data in which there was a specified marginal hazard ratio. To do so, we modified previously described data-generating processes for generating data with a specified marginal odds ratio or risk difference [23, 24]. We used an iterative process to determine the value of β_{treat} (the conditional log-hazard ratio) that induced the desired marginal hazard ratio. Briefly, using the aforementioned conditional model, we simulated a time-to-event outcome for each subject, first assuming that the subject was untreated and then assuming that the subject was treated. In the sample consisting of both potential outcomes (survival or event time under lack of treatment and survival or event time under treatment), we regressed the survival outcome on an indicator variable denoting treatment status. The coefficient for the treated status indicator denotes the log of the marginal hazard ratio. We repeated this process 1000 times to obtain an estimate of the log of the marginal hazard ratio associated with a specific value of β_{treat} in our conditional outcome model. We then employed a bisection approach to determine the value of β_{treat} that resulted in the desired marginal hazard ratio. We applied this process twice: first to determine the value of β_{treat} that induced a desired marginal hazard ratio in the overall population. We will describe this as the ATE in the population. Second, we repeated the process but used only subjects who were ultimately assigned to the treatment when estimating the marginal hazard ratio (i.e., we fit the Cox model on the dataset of potential outcomes restricted to those subjects who were ultimately treated). We thus determined the value of β_{treat} that induced a desired marginal hazard ratio in the treated population. We describe this as the ATT. We acknowledge that describing these hazard ratios as average treatment effects is a slight abuse of convention as they do not explicitly involve taking expectations of differences in potential outcomes. However, our intent was to describe the average effect in the entire population or in the population of treated subjects.

We allowed the following factors to vary in our Monte Carlo simulations: the percentage of subjects that were treated (5%, 10%, and 25%) and the true marginal hazard ratio (0.8, 1, 1.10, 1.25, 1.50, 1.75, and 2). We thus examined 21 scenarios (three treatment prevalences × seven marginal hazard ratios). For each true marginal hazard ratio, we considered both the ATE and ATT hazard ratios. In each scenario, we simulated 10,000 datasets, each consisting of 10,000 subjects.

Within each simulated dataset, we did the following: we estimated the propensity score using a logistic regression model to regress treatment status on the seven baseline covariates that affected the outcome. We selected this approach to variable selection for the propensity score model, as it has been shown to result in better estimation compared with selecting only those variables that affect treatment selection [25].

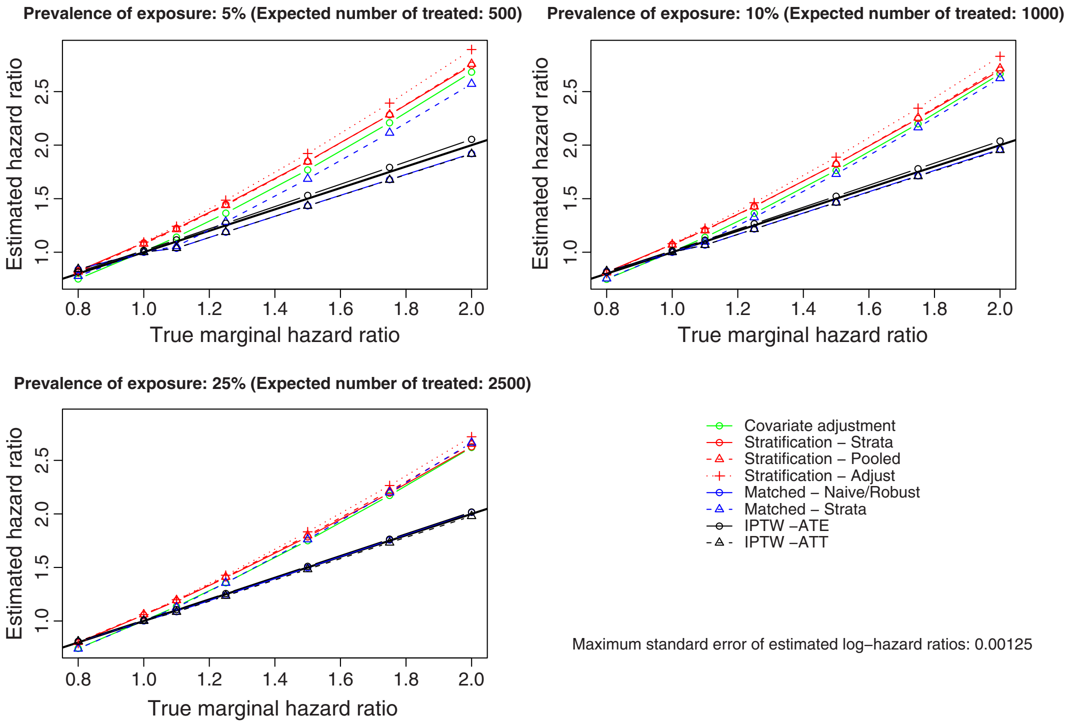
In each of the 10,000 simulated datasets for each scenario, we estimated the log-hazard ratio and its standard error using the methods described in Section 2. Let θ_i denote the estimated log-hazard ratio obtained from the ith simulated dataset using a given method, whereas θ denotes the true log-marginal hazard ratio. We estimated the mean treatment effect (on the log-hazard scale), bias, and mean squared error (MSE) as $\bar{\theta} = (1/10000)\sum_{i=1}^{10000} \theta_i$, $\text{Bias} = (1/10000)\sum_{i=1}^{10000} (\theta_i - \theta)$, and $\text{MSE} = (1/10000)\sum_{i=1}^{10000} (\theta_i - \theta)^2$, respectively. We defined relative bias as $100 \times (\text{Bias}/\theta)$. We also examined the accuracy with which the estimated standard error of the estimated log-hazard ratio estimated the sampling variability of the estimated log-hazard ratio. To do so, we compared two estimates. First, within each of the 10,000 simulated datasets, we estimated the standard error of estimated log-hazard ratio; we then determined the mean standard error of the log-hazard ratio across the 10,000 simulated datasets. Second, we determined the standard deviation of the estimated log-hazard ratios across the 10,000 simulated datasets. The first quantity estimates the mean standard error, whereas the second quantity estimates the sampling variability of the log-hazard ratio. We then determined the ratio of these two quantities. If the ratio equals 1, then the estimated standard error of the log-hazard ratio is correctly estimating the sampling variability of the estimated log-hazard ratio. Finally, within each simulated dataset and for each method, we computed the 95% confidence interval for the estimated hazard ratio. We then determined the mean length of the estimated 95% confidence intervals as well as the proportion of 95% confidence intervals that covered the true hazard ratio that was used in the data-generating process.

3.2. *Monte Carlo simulations—results*

When the proportion of subjects who were treated was 0.05, 0.10, and 0.25, then the average number of matched pairs formed across the 10,000 simulated samples for each scenario was 499.9, 993.4, and 2357.4, respectively. Thus, we matched approximately 100%, 99.3%, and 94.3% of treated subjects to an untreated subject. Thus, the matched methods should have minimal bias due to incomplete matching [13].

Figure 1 shows the exponential of the mean estimated treatment effect $\left(\exp\left(\bar{\theta}\right)\right)$, whereas Figure 2 shows the relative bias in estimating the log-marginal-hazard ratios (note that the relative biases are not reported for the scenarios with a true hazard ratio of 1 as the log-hazard ratio is 0). In Figure 1, we have added a solid diagonal line of unit slope. Deviation from this diagonal line indicates biased estimation of marginal hazard ratios. The use of 10,000 simulated datasets resulted in precise estimation of the log-hazard ratio. Across the 21 scenarios and the different estimation methods, the maximum standard error of the estimated log-hazard ratio was 0.00125. Both bias and relative bias tended to be substantial for covariate adjustment using the propensity score and for the three stratification methods. When we use matching, the naïve and robust methods resulted in the same relative bias as they are both marginal models of the same functional form and resulted in the same estimate of the regression coefficient for the log-hazard ratio. The method that stratified on matched pairs resulted in substantially greater bias than did the other two matching methods. IPTW, using either set of weights, and two of the matching methods (naïve and robust) resulted in estimates with minimal bias.

The preceding results identify several issues that must be addressed before proceeding to examination of further results. The naïve Cox model and the robust Cox model in the propensity-score-matched sample are marginal models—they estimate population-average effects. Similarly, the IPTW method



Maximum standard error of estimated log–hazard ratios: 0.00125

Figure 1. Estimated hazard ratio using different propensity score methods. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

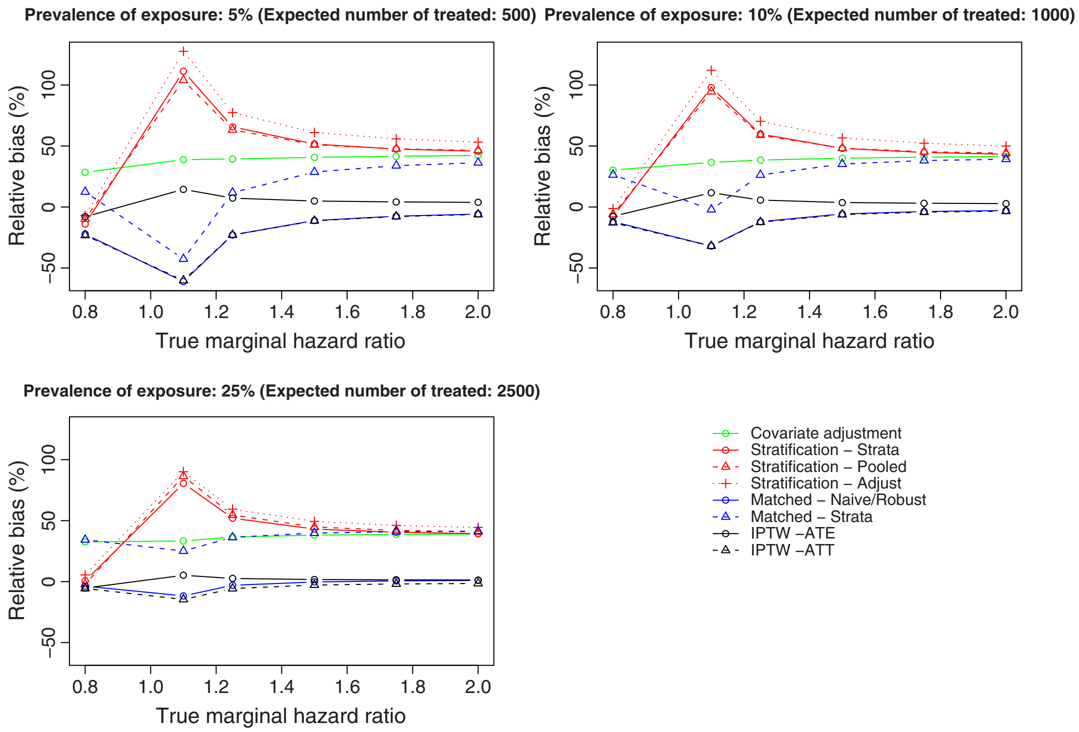


Figure 2. Relative bias of different propensity score methods for estimating marginal hazard ratios. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

estimates marginal effects. The high biases observed for covariate adjustment using the propensity score, stratification using the propensity score, and the Cox model that stratified on matched pairs in the propensity-score-matched sample highlight that these methods do not estimate marginal effects. Instead, they result in conditional measures of effect—they estimate conditional hazard ratios. That this is true for covariate adjustment using the propensity score is obvious, and it did not require simulations to illustrate this fact. That it is true for the different stratified approaches may be less obvious. With the stratification (adjusted) approach, one is fitting a conditional model that estimates the effect of treatment selection after *adjusting* for the propensity score strata (as a categorical variable). The stratification (strata) approach fits a univariate Cox model that stratifies on the propensity score strata, thereby allowing the baseline hazard function to vary across the propensity score strata. However, one is still deriving estimates of treatment effect from a model that conditions on the strata. The stratification (pooled) approach fits stratum-specific univariate Cox models. However, one then pools the stratum-specific hazard ratios. Pooling or averaging these stratum-specific estimates results in a smoothed estimate that represents an adjusted or conditional effect. Thus, none of these approaches allows one to estimate a marginal hazard ratio. Similarly, the model fit in the propensity-score-matched sample that stratified on the matched pairs is a conditional model that conditions on the matched pairs. Thus, none of these methods results in marginal estimates of treatment effect.

Our data-generating process employed a conditional Cox model to generate outcomes. We selected the conditional log-hazard ratio so as to induce a specified marginal hazard ratio. We examined the relative bias of covariate adjustment using the propensity score, stratification on the propensity score, and the matched analysis that stratified on matched pairs when estimating the underlying conditional hazard ratio that was used in the data-generating process. Because this set of propensity score methods resulted in biased estimation of the underlying marginal hazard ratio, it is important to examine whether this set of methods also result in biased estimation of the conditional hazard ratio used in the data-generating process. Figure 3 shows the relative bias for estimating the conditional hazard ratio. Each method resulted in moderate bias in estimating the conditional hazard ratio used in the data-generating process. In theory, there are multiple conditional effects. Indeed, there is potentially a different conditional effect for every set of covariates for which one adjusts in a regression model. Thus, although these methods allow

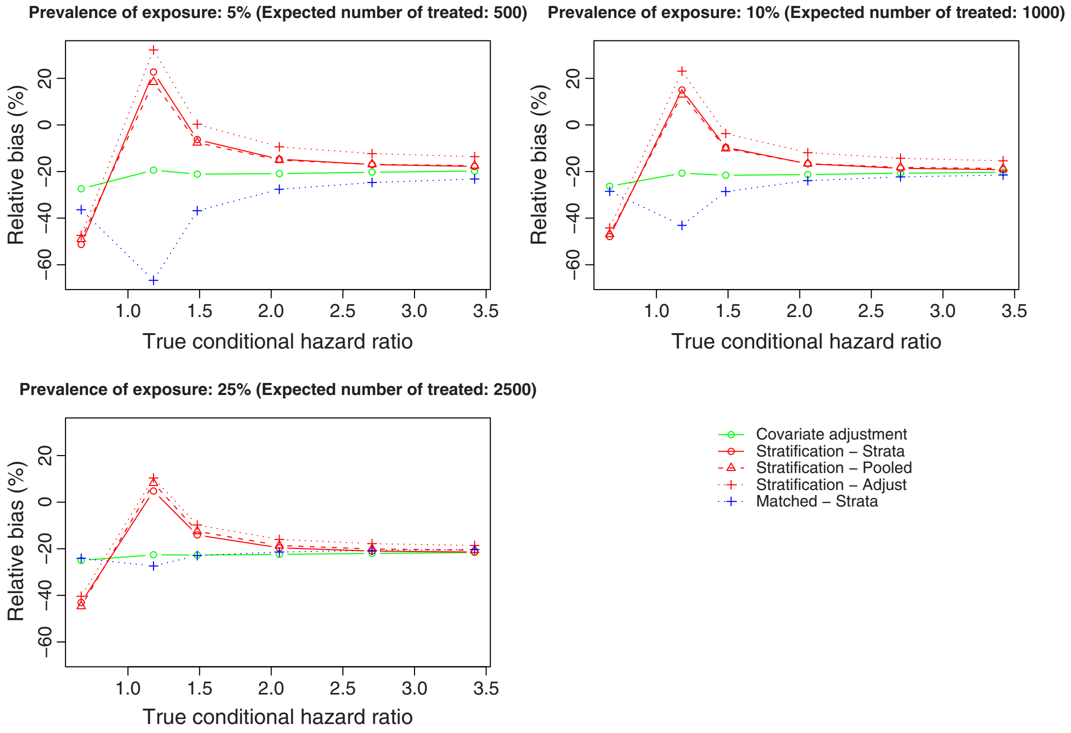


Figure 3. Relative bias of different propensity score methods for estimating conditional hazard ratios.

estimation of conditional effects, it is unclear which conditional effect they are estimating. In particular, the conditional effect that is being estimated does not appear to coincide with the conditional effect employed in the data-generating process.

For the reasons noted earlier, for the remainder of this section, we focus on the performance of the two IPTW methods (ATE and ATT weights), the naïve Cox model in the propensity-score-matched sample, and the robust Cox model in the propensity-score-matched sample for estimating marginal hazard ratios.

In Figure 4, we report, for each estimation method, the ratio of the mean standard error of the estimated log-hazard ratio across the 10,000 simulated datasets (i.e., in each simulated dataset, we obtained an estimated standard error of the estimated treatment effect; we then averaged these across the 10,000 simulated datasets) to the standard deviation of the estimated log-hazard ratios across the 10,000 simulated datasets (i.e., in each dataset, we obtained an estimated log-hazard ratio; we then estimated the sampling variability of these estimated log-hazard ratio). This analysis indicates whether, for a given estimation method, the estimated standard error of the estimated treatment effect is correctly estimating the sampling variability of the estimated treatment. This ratio tended to be larger for the naïve Cox model in the matched sample than for the robust Cox model in the matched sample. The mean ratio for the naïve matched method was 1.28 across the 21 scenarios, whereas it was 1.09 for the robust matched method. Thus, ignoring the matched nature of the propensity-score-matched sample resulted in estimates of the standard error of the log-hazard ratio that was inflated by an average of 28% (i.e., the naïve matched method resulted in estimates of standard error that were, on average, 28% too large, whereas the robust matched approach resulted in estimates of standard error that were, on average, 9% too large). The mean ratios for the IPTW methods were 1.0 and 1.40 when using ATE and ATT weights, respectively.

Figure 5 shows coverage rates of 95% confidence intervals. Because of our use of 10,000 iterations per scenario in our Monte Carlo simulations, any confidence intervals whose empirical coverage rate is less than 0.9457 or greater than 0.9543 would be statistically significantly different from 0.95, using a standard test based on the normal approximation to the binomial distribution. In general, coverage rates of 95% confidence intervals were suboptimal. Coverage rates of 95% confidence intervals tended to improve as the proportion of subjects who were treated increased. Figure 6 shows the mean length of the estimated 95% confidence intervals. Among the three methods that estimated the ATT, the naïve matched approach tended to result in estimated 95% confidence intervals that were slightly wider than those from the other two approaches. The differences between these three methods decreased as the proportion of subjects who were treated increased.

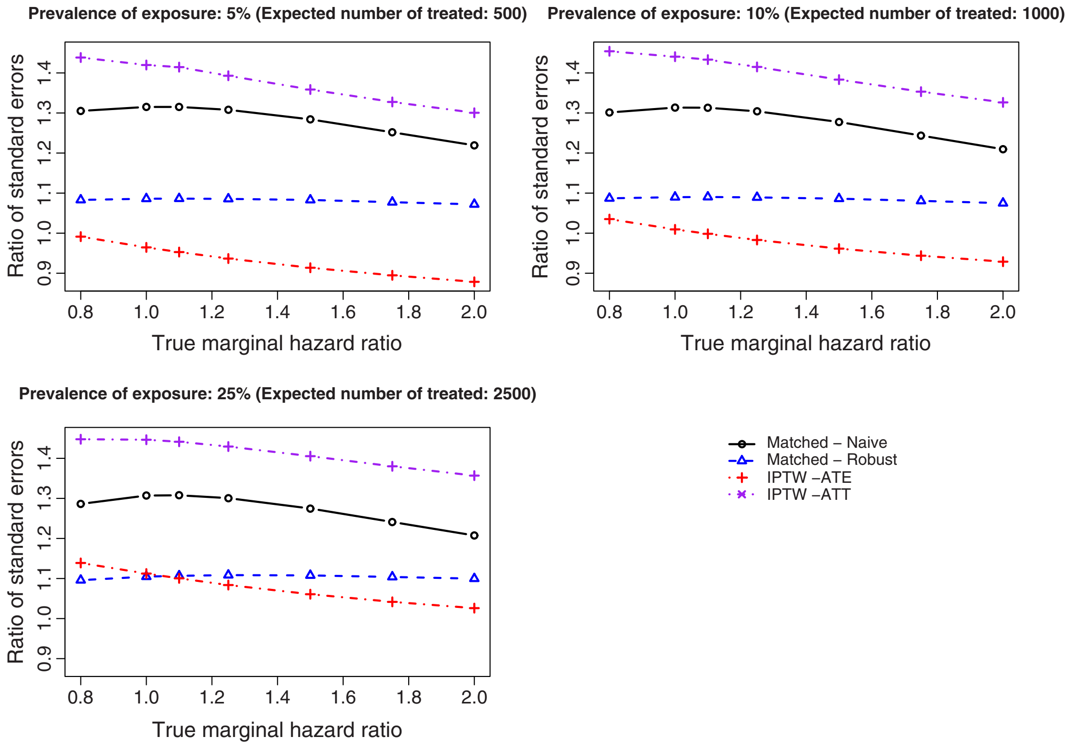


Figure 4. Ratio of mean standard error to standard deviation of estimated log-hazard ratios. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

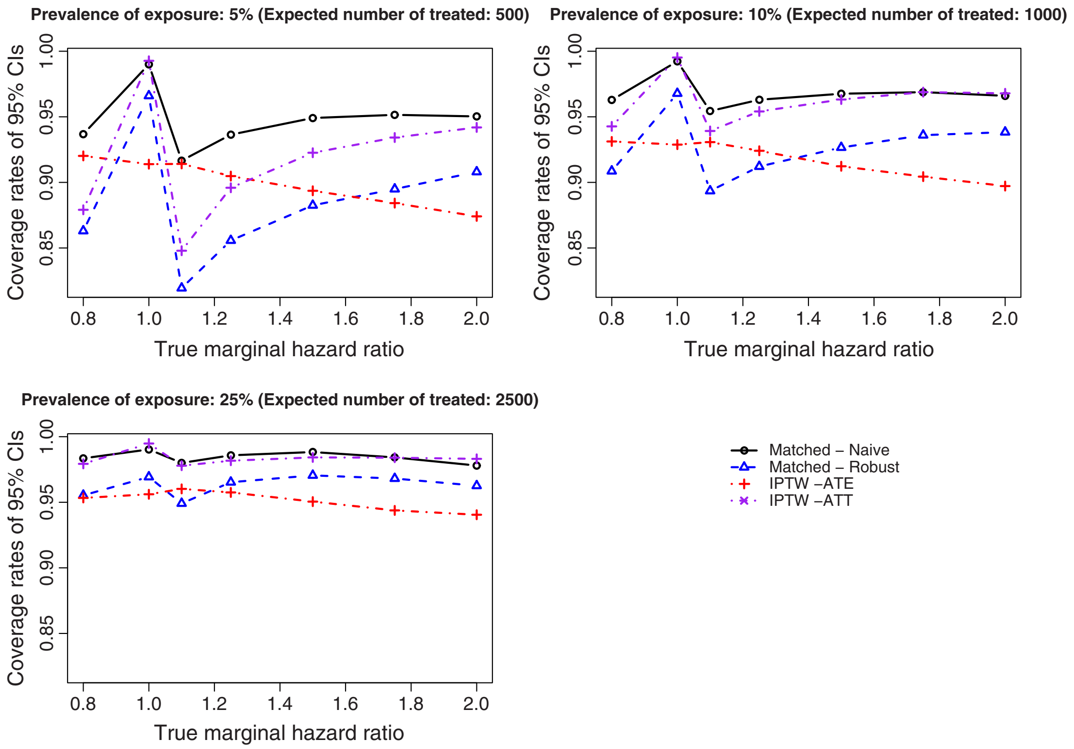


Figure 5. Coverage rates of 95% confidence intervals (CIs) for estimated hazard ratios. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

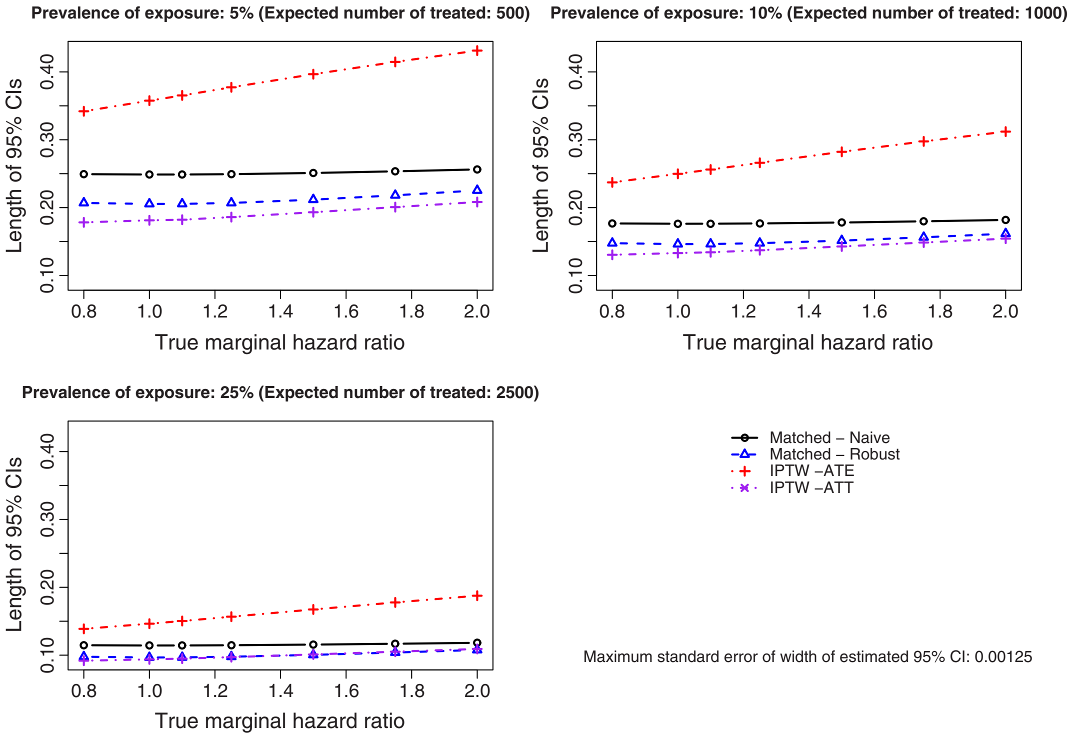


Figure 6. Mean length of 95% confidence intervals (CIs) for marginal hazard ratios. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

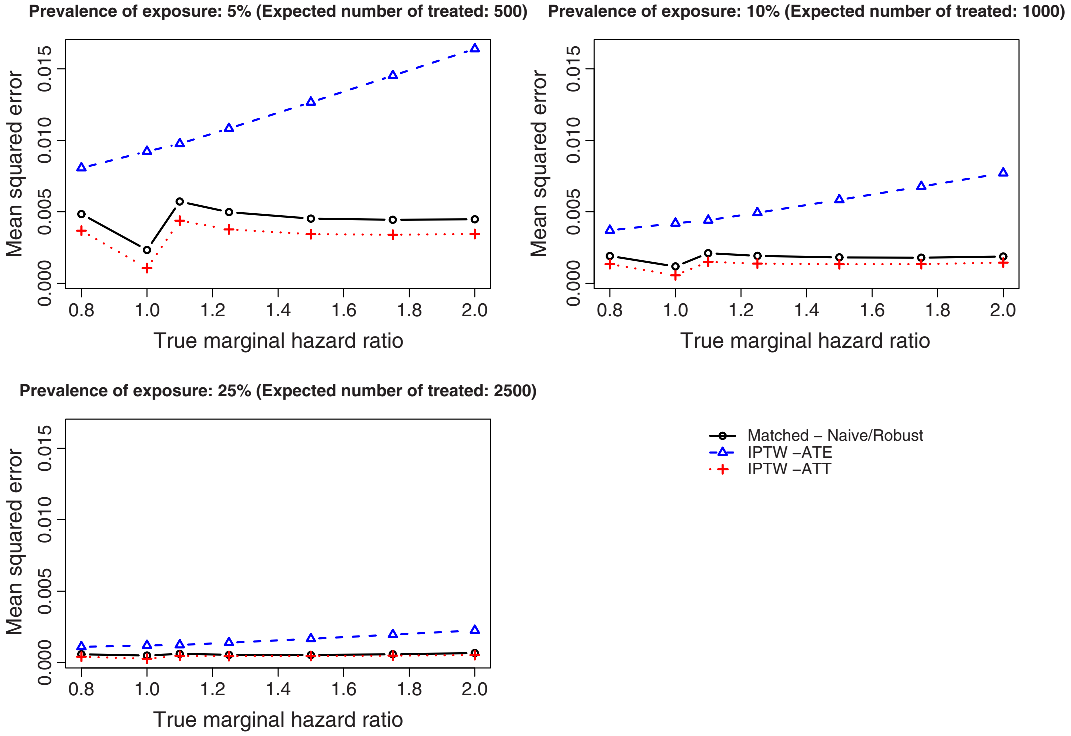


Figure 7. Mean squared error of estimated log-hazard ratio. IPTW, inverse probability of treatment weight; ATE, average treatment effect; ATT, average treatment effect in the treated.

Figure 7 presents the MSE of the estimated treatment effects. IPTW using the ATE weights is the only method that allows one to estimate an ATE effect. Therefore, we do not compare the performance of this method with that of the other methods, as the other methods are estimating a different quantity (the marginal effect of treatment in the treated population). However, we note that with the IPTW-ATE approach, the MSE increased as the true marginal hazard ratio increased. Because the naïve Cox model and the robust Cox model fits in the propensity-score-matched sample produce the same point estimate of the log-hazard ratio, these two methods result in estimates with the same MSE. In all 21 of the scenarios, the IPTW (ATT weights) estimator had lower MSE than did the matched estimator. Across the 21 scenarios, the mean MSE of the matched estimates was 0.0023, whereas it was 0.0017 for the IPTW (ATT) estimates. However, differences between these two approaches decreased as the proportion of subjects that were treated increased.

4. Discussion

We conducted an extensive series of Monte Carlo simulations to examine the performance of different propensity score methods to estimate marginal hazard ratios. We briefly summarize our findings and place them in the context of the existing literature.

Greenland and others distinguished between measures of effect that are collapsible and those that are non-collapsible [10]. A measure of effect is collapsible if, in the absence of confounding, the conditional effect and the marginal effect coincide. Differences in means and risk differences are collapsible, whereas odds ratios and hazard ratios are non-collapsible [10, 11]. Neuhaus *et al*. demonstrated that, in general, the marginal odds ratio will be closer to the null than the conditional effect [26]. Similarly, Pocock *et al*. commented that in RCTs, the adjusted (i.e., conditional) odds ratio or hazard ratio will be further from the null than the unadjusted (e.g., marginal) estimate [27]. Stratification on the propensity score and covariate adjustment using the propensity score reduce bias when estimating linear treatment effects [2], which are collapsible. The current study found that these methods do not perform well when outcomes are time to event in nature and when the hazard ratio is used as the measure of effect. Both of these methods resulted in biased estimation of the marginal hazard ratio. Furthermore, they also resulted in biased estimation of the conditional hazard ratio that underlay the data-generating process. These two methods are likely estimating a conditional effect; however, it is unclear which conditional effect it is. It does not appear to be the conditional effect that would be obtained by specifying the outcome regression model correctly and including the covariates that are predictive of the outcome.

When outcomes are time to event in nature, we can summarize the effect of treatment in at least three ways: first, the hazard ratio describing the relative effect of the treatment on the hazard of the outcome; second, the estimation of Kaplan–Meier survival curves in each treatment group, which allows for estimating the absolute effect of treatment on the probability of the event occurring within any specified duration of follow-up; third, mean or expected survival can be estimated in each treatment group, and the effect of treatment on mean survival can be reported. Of these three methods, the first provides a relative measure of effect, whereas the latter two provide absolute measures of effect. The first two are the most common in reports of RCTs, whereas the latter is infrequently used in RCTs [5]. In the current study, we have focused on the estimation of hazard ratios, in part, because of the frequency with which they are reported in the medical literature. There are different ways in which survival curves can be estimated and compared using propensity score methods. In a propensity-score-matched sample, we can estimate Kaplan–Meier estimates of survival curves in each treatment group separately. We have removed confounding by design: the distribution of baseline covariates will be similar between treated and untreated subjects in the matched sample. Thus, the use of the ‘crude’ Kaplan–Meier estimator can allow for an unbiased comparison of survival between treatment groups. Because of the lack of independence between the two matched samples, the log-rank test should not be used for testing the equality of the survival curves [28]. Both Cole and Hernan [29] and Xie and Liu [30] have described methods to estimate survival curves in the sample weighted by the inverse of the probability of treatment. The performance of different propensity score methods for estimating the latter two measures of effect should be examined in subsequent research.

There are certain limitations to the current study. Our findings were based on an extensive series of Monte Carlo simulations. As such, our findings warrant replication in different scenarios and under different assumptions about the distribution of baseline covariates and about the number of measured baseline covariates and their relationship with treatment selection and with outcome. However, given our focus on estimating marginal hazard ratios, analytic determination of the performance of estimation would be very difficult, particularly for the methods based on propensity score matching. Furthermore, we would note that several prior studies examining the performance of propensity score methods for estimating treatment effects have employed Monte Carlo simulations [6–9, 31, 32]. We speculate that our findings will generalize to other combinations of baseline covariates. A second limitation relates to our focus on 1:1 matching on the propensity score, in which pairs of treated and untreated subjects were formed. We did not consider M:1 matching in which M untreated subjects are matched to each treated subject. M:1 matching was examined in a prior publication, in which it was found that, in many scenarios, use of $M = 1$ or 2 tended to be optimal [33]. Increasing the number of untreated subjects matched to each treated subject will, on average, tend to result in the matching of increasingly dissimilar subjects. This can result in increased bias in the estimated treatment. However, there will tend to be a commensurate increase in precision of the estimated treatment effect. In the current study, we found that the robust matching approach and the IPTW (ATT weights) tended to result in estimates with comparable bias. We speculate that matching multiple untreated subjects to each treated subjects would, on average, result in increased bias compared with the weighted estimator. A third limitation is our inclusion of only a single matching algorithm: greedy nearest-neighbor matching on the logit of the propensity score using calipers defined by the variance of the logit of the propensity score. This approach was included as it has been found to perform well compared with other commonly used alternatives [34]. However, we did not examine other approaches such as optimal matching because of its increased computational complexity [35]. It is possible that the use of optimal matching may result in improved performance of the matching approach compared with the other propensity score approaches.

The findings of the current study complement those of previously published studies. Our examination of the estimation of marginal hazard ratios using propensity score matching confirmed the observation made in prior studies that variance estimation should account for the matched nature of the sample [28, 36]. In particular, the use of a robust variance estimator resulted in estimates of standard error that tended to better reflected the sampling distribution of the estimated log-hazard ratio than did the use of naïve model-based standard errors from a maximum likelihood model that assumed independent observations. In a recently published study, Gayat *et al*. examined the performance of different implementations of propensity score matching and covariate adjustment using the propensity score for estimating hazard ratios [37]. When using propensity-score matching, they examined a variety of estimators in the propensity-score matched sample: a crude (unadjusted) estimator, a crude (stratified) estimator, and an estimator that was adjusted for all of the measured covariates. Similar to the current study, they found that, in the matched sample, it was preferable to use a robust variance estimator, rather than the naïve variance estimator. The current study examined IPTW methods and stratification on the propensity score, which were not considered in the study by Gayat *et al*. The conclusions of these two studies differed in one respect: they found that the stratified matched approach and covariate adjustment using the propensity score resulted in at most modest bias when estimating *conditional* hazard ratios (i.e., relative biases of less than 10%), whereas we found larger biases when estimating conditional hazard ratios. We speculate that are several potential explanations for these differences: first, in the current simulations, covariates were generated from normal distributions, whereas in the paper by Gayat *et al*., they were simulated from Bernoulli distributions. Second, while the earlier paper only considered three hazard ratios, we considered seven hazard ratios in the current study. The fact that we observed moderate biases when estimating conditional hazard ratios indicates that, in general, propensity score methods result in biased estimation of the underlying conditional hazard ratio. However, the degree of the bias will vary across different settings.

The finding that propensity-score methods result in confidence intervals whose empirical coverage rates differ from the advertised levels warrants concern. This finding is similar to observations made in an earlier study comparing different propensity score methods for estimating absolute risk reductions or differences in proportions [9]. We speculate that the sub-optimal coverage rates are related to estimation of the standard error of the estimated log-hazard ratio, rather than to bias in the estimated log-hazard ratio. As illustrated in Figure 1, matching on the propensity score and IPTW result in at most minimal bias in the estimation of the true log-hazard ratio. However, as illustrated in Figure 4, both methods result in mis-estimation of the variance of the sampling distribution of the estimated log-hazard ratio. Since the confidence interval is constructed using the estimated log-hazard ratio and the estimated standard error of the log-hazard ratio, it appears likely that the sub-optimal coverage rates relates to the mis-estimation of the standard error of the log-hazard ratio. Reasons for why this occur merit further investigation. One avenue that warrants exploration is the use of estimated versus known propensity scores. Propensity score methods rely on estimation of the propensity score. However, once the propensity score is estimated, the analyst acts as though it was a known quantity, rather than an estimate of an unknown quantity. Future research is required on methods for variance estimation that account for the fact that the propensity score has been estimated.

In summary, researchers should employ propensity score matching and inverse probability of treatment weighting using the propensity score when estimating the relative effect of treatment on time-to-event outcomes using observational or non-randomized data. Use of these methods allows for estimation of the marginal effect of treatment on survival—the same metric that is reported in reports of RCTs with time-to-event outcomes. An advantage to the use of IPTW using the propensity score is that by using different weights, one can estimate both ATE and ATT. While both matching and IPTW resulted in estimates of marginal hazard ratios with minimal bias, the latter approach resulted in estimates with lower MSE, suggesting that the estimates resulting from weighted analyses have improved precision than those resulting from matched analyses.

Acknowledgements

The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported by an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation.

References

1. Cook TD, DeMets DL. *Introduction to Statistical Methods for Clinical Trials*. Chapman & Hall/CRC: Boca Raton, FL, 2008.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
3. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and Drug Safety* 2004; **13**(12):841–853.
4. 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.
5. 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.
6. 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.
7. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; **26**(16):3078–3094.
8. Austin PC. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology* 2008; **61**(6):537–545.
9. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine* 2010; **29**(20):2137–2148.
10. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology* 1987; **125**(5):761–768.
11. 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.
12. Rosenbaum PR. Propensity score. In *Encyclopedia of Biostatistics*, Armitage P, Colton T (eds). John Wiley & Sons: Boston, 2005; 4267–4272.
13. 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.
14. 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.
15. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *Journal of the American Statistical Association* 1989; **84**:1074–1078.
16. Cummings P, McKnight B, Greenland S. Matched cohort methods for injury research. *Epidemiologic Reviews* 2003; **25**:43–50.
17. 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.
18. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**(2):295–313.
19. 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.
20. Morgan SL, Todd JL. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology* 2008; **38**:231–281.
21. 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.

22. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723.
23. Austin PC. A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics - Simulation and Computation* 2010; **39**:563–577. DOI: 10.1080/03610910903528301.
24. Austin PC, Stafford J. The performance of two data-generation processes for data with specified marginal treatment odds ratios. *Communications in Statistics - Simulation and Computation* 2008; **37**:1039–1051.
25. 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.
26. 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.
27. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**(19):2917–2930.
28. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *International Journal of Biostatistics* 2009; **5**: Article 13. DOI: 10.2202/1557-4679.1146.
29. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 2004; **75**:45–49.
30. 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.
31. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**(3):337–346.
32. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiolgy and Drug Safety* 2008; **17**(6):546–555.
33. Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *American Journal of Epidemiology* 2010; **172**(9):1092–1097.
34. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical Journal* 2009; **51**(1):171–184.
35. Rosenbaum PR. *Observational Studies*. Springer-Verlag: New York, NY, 2002.
36. 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.
37. 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.