Statistics
in Medicine

Research Article

Received 21 March 2014, Accepted 7 July 2014 Published online 4 August 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6276

The use of bootstrapping when using propensity-score matching without replacement: a simulation study

Peter C. Austin[a,b,c*†] **and Dylan S. Small**[d]

Propensity-score matching is frequently used to estimate the effect of treatments, exposures, and interventions when using observational data. An important issue when using propensity-score matching is how to estimate the standard error of the estimated treatment effect. Accurate variance estimation permits construction of confidence intervals that have the advertised coverage rates and tests of statistical significance that have the correct type I error rates. There is disagreement in the literature as to how standard errors should be estimated. The bootstrap is a commonly used resampling method that permits estimation of the sampling variability of estimated parameters. Bootstrap methods are rarely used in conjunction with propensity-score matching. We propose two different bootstrap methods for use when using propensity-score matching *without replacement* and examined their performance with a series of Monte Carlo simulations. The first method involved drawing bootstrap samples from the matched pairs in the propensity-score-matched sample. The second method involved drawing bootstrap samples from the original sample and estimating the propensity score separately in each bootstrap sample and creating a matched sample within each of these bootstrap samples. The former approach was found to result in estimates of the standard error that were closer to the empirical standard deviation of the sampling distribution of estimated effects. © 2014 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd.

Keywords: propensity score; propensity-score matching; bootstrap; variance estimation; Monte Carlo simulations; matching

1. Introduction

Propensity-score methods are increasingly being used to reduce or minimize the confounding that occurs frequently in observational studies. The propensity score is the probability of treatment assignment conditional on measured baseline covariates (i.e., pre-treatment covariates that are measured prior to the application of the treatment) [1]. Matching on the propensity score is a commonly used method for removing the effects of confounding due to observed covariates [2–4]. Matching on the propensity score entails forming matched sets of treated and untreated subjects who have a similar value of the propensity score [5]. The most common implementation of propensity-score matching is pair matching, in which pairs of treated and untreated subjects are formed. Rosenbaum and Rubin demonstrated that subjects who have the same propensity score will have the same distribution of measured baseline covariates [1]. A consequence of this is that matched treated and untreated subjects will have measured baseline covariates that are more likely to be similar to one another than are the baseline covariates of two unmatched subjects. Therefore, it is likely that a within-matched pair correlation of outcomes has been induced by matching on the propensity score. There is some controversy surrounding variance estimation when estimating

[a]*Institute for Clinical Evaluative Sciences, Toronto, Canada*
[b]*Institute of Health Management, Policy and Evaluation, University of Toronto, Toronto, Canada*
[c]*Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Canada*
[d]*Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA, U.S.A.*
[*]*Correspondence to: Peter C. Austin, Institute for Clinical Evaluative Sciences, G106, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada.*
[†]*E-mail: peter.austin@ices.on.ca*

treatment effects in propensity-score-matched samples. Schafer and Kang have suggested that methods of inference appropriate for independent samples can be used for assessing the statistical significance of treatment effects when using propensity-score matching [6]. In contrast to this, other authors, using Monte Carlo simulations, have found that statistical methods that account for the matched nature of the sample better approximate the sampling distribution of the estimated treatment effect [7–10].

The bootstrap is a well-known, resampling method for estimating the standard error of estimated statistics and of constructing confidence intervals [11]. Despite its widespread use in statistics, its use in conjunction with propensity-score methods is rare. Given the debate surrounding appropriate methods for estimating the sampling variability of effects estimated using propensity-score matching, bootstrap methods may hold promise for use with propensity-score matching.

The objective of the current study was to examine the performance of bootstrap methods to estimate the sampling variability of treatment effect estimates obtained using propensity-score matching. The paper is structured as follows. In Section 2, we describe three different propensity-score matching algorithms and two different bootstrap approaches that could be used with propensity-score matching. In Section 3, we describe the design of an extensive series of Monte Carlo simulations to assess the performance of bootstrap methods to estimate the sampling variability of estimated treatment effects. In Section 4, we report the results of these simulations. Finally, in Section 5, we summarize our findings and place them in the context of the existing literature.

2. Background: statistical methods for matching and bootstrapping

In this section, we describe three algorithms for matching on the propensity score and two ways in which the bootstrap can be used in conjunction with propensity-score matching to estimate the standard error of the estimated effect of treatment. Each of the matching algorithms uses matching *without* replacement, so that each untreated or control subject is included in at most one matched set.

2.1. Three algorithms for 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 [5]. 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. The first algorithm that we examined was greedy nearest-neighbor matching (NNM) on the propensity score, with a random ordering of treated subjects [6]. Using this approach, a treated subject is selected at random. This treated subject is then matched to the untreated subject whose propensity score is closest to that of the treated subject. Matching *without* replacement was used, so that once an untreated subject was selected for matching to a given treated subject, that untreated subject was no longer eligible for consideration as a potential match for subsequent treated subjects. Thus, each untreated subject was included in at most one matched set.

The second matching algorithm that we used was greedy NNM on the logit of the propensity score within specified caliper widths to form pairs of treated and untreated subjects [5]. 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 [12]. This algorithm is a refinement of NNM, in which an untreated subject is only considered as a potential match for a given treated subject if the difference in the logits of their propensity scores is below some maximal difference (the caliper width). As above, we considered matching *without* replacement, so that once an untreated subject was matched to a given treated subject, that untreated subject was no longer eligible for consideration as a potential match for subsequent treated subjects.

The third matching algorithm that we used was optimal matching [13]. Optimal matching forms matched pairs of treated and untreated subjects so as to minimize the average within-pair difference in the propensity score.

2.2. Estimating treatment effects in propensity-score-matched samples

In the current study, we considered three different types of outcomes: continuous outcomes, binary outcomes, and survival or time-to-event outcomes. For continuous outcomes, the effect of treatment can be estimated as the difference between the mean outcome in the treated subjects in the matched sample and the mean outcome in the untreated subjects in the matched sample. This is equivalent to the mean within-pair difference in the outcome in the matched sample. The standard error of the estimated difference in

means can be estimated as the standard error of the within-pair differences in the outcome. When outcomes are binary, the risk difference or absolute risk reduction can be estimated similarly. The variance of the risk difference can be estimated using methods that account for the matched nature of the sample [9, 14]. When outcomes are time-to-event in nature, a Cox proportional hazards regression model can be used to regress survival on an indicator variable denoting treatment status. Prior research has found that the use of a robust, sandwich-type estimate of the variance of the regression coefficient that accounts for the clustering within matched sets [15] performs better than using the naïve model-based standard errors [10].

2.3. *Bootstrap methods for use with propensity-score matching*

A bootstrap sample is a sample drawn using sampling with replacement from the original sample, such that the size of the bootstrap sample is equal to that of the original sample [11]. We describe two different methods for using the bootstrap to estimate the sampling variability of the estimated treatment effect using a propensity-score-matched sample. We refer to these two approaches as the simple bootstrap and the complex bootstrap. We also describe three different bootstrap methods for estimating 95% confidence intervals with either the simple bootstrap or the complex bootstrap.

2.3.1. The simple bootstrap. The simple bootstrap is simply the conventional bootstrap applied to the matched sample that was obtained using propensity-score matching (using whatever matching algorithm was selected). The only caveat is that one must bootstrap matched pairs, rather than individual subjects. Let the matched sample consist of N matched pairs: $M_1, M_2, \ldots, M_N$, where each matched pair consists of a treated subject and an untreated subject who have similar values of the propensity score. B bootstrap samples are then drawn from the set of N matched pairs: $\mathrm{A} = \{M_1, M_2, \ldots, M_N\}$. Thus, each bootstrap sample consists of N matched pairs, drawn with the replacement from the set A of matched pairs. In each of these B bootstrap samples, the effect of treatment is estimated using the appropriate method described in Section 2.2. The standard deviation of the estimated treatment effects across the B bootstrap samples is used as an estimate of the standard error of the estimated treatment effect in the original propensity-score-matched sample.

Confidence intervals can be estimated using three different approaches. First, one can use a standard normal-theory approach. Having estimated the standard error of the estimated treatment effect, one can use $\hat{\theta} \pm 1.96\mathrm{se}(\hat{\theta})$ as the endpoints of the 95% confidence interval, where $\hat{\theta}$ denotes the estimated effect of treatment in the original matched sample, while $\mathrm{se}(\hat{\theta})$ denotes the bootstrap estimate of the standard error of the estimated treatment effect. Second, one can use a nonparametric percentile-based approach to estimate 95% confidence intervals. To do so, one uses the 2.5th and 97.5th percentiles of the estimated treatment effects across the B bootstrap samples. Third, one can construct accelerated and bias-corrected ($\mathrm{BC_a}$) confidence intervals [11].

2.3.2. The complex bootstrap. The complex bootstrap attempts to incorporate additional potential sources of variability when estimating the sampling variability of the estimated treatment effect. In particular, it attempts to incorporate two additional sources of variability compared with that addressed by the simple bootstrap: variability in estimating the propensity-score model and variability in the formation of the propensity-score-matched sample. Using this approach, B bootstrap sample are drawn from the original (unmatched) sample. In each of the B bootstrap samples, the propensity-score model is estimated, and a propensity-score-matched sample is formed using the selected matching algorithm. Then, in each of the B matched samples, the effect of treatment is estimated using the appropriate method described in Section 2.2. The standard deviation of the estimated treatment effects across the B propensity-score-matched samples is then estimated. This serves as an estimate of the sampling distribution of the estimated treatment effect in the original propensity-score-matched sample. The three methods described earlier can be used to estimate 95% confidence intervals when using the complex bootstrap.

It should be noted that while the complex bootstrap accounts for more sources of potential variability, it is also substantially more computationally intensive. This is because it requires B additional implementations of the propensity-score matching algorithm, whereas the simple bootstrap method only involves drawing bootstrap samples from the original propensity-score-matched sample. In general, matching algorithms are more computationally intensive than are random sampling algorithms.

3. Monte Carlo simulations: methods

We used a series of Monte Carlo simulations to examine performance of two different bootstrap algorithms for estimating the sampling variability of treatment effect estimates obtained using propensity-score matching. We considered continuous, binary, and survival outcomes.

3.1. Data-generating processes

We simulated data for a setting in which there were 10 baseline covariates (X_1–X_{10}). These covariates were simulated from independent standard normal distributions. Of these 10 covariates, seven affected treatment selection (X_1–X_7), while seven affected the outcome (X_4–X_{10}). Furthermore, covariates were allowed to have a weak, moderate, strong, or very strong effect on treatment selection or outcome.

For each subject, the probability of treatment selection was determined from the following logistic model: $\text{logit}(p_i) = \alpha_{0,\text{treat}} + \alpha_W x_1 + \alpha_M x_2 + \alpha_S x_3 + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS} x_7$. The intercept of the treatment-selection model ($\alpha_{0,\text{treat}}$) was selected so that the proportion of subjects in the simulated sample that were treated was fixed at the desired proportion (5%, 10%, 20% or 25% of subjects were exposed to the treatment). 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-selection effects. For each subject, treatment status was generated from a Bernoulli distribution with subject-specific parameter $p_i : Z_i \sim Be(p_i)$.

We generated a continuous outcome, a binary outcome, and a time-to-event outcome for each subject. The continuous outcome was generated as $Y_i = Z_i + \alpha_W x_4 + \alpha_M x_5 + \alpha_S x_6 + \alpha_{VS} x_7 + \alpha_W x_8 + \alpha_M x_9 + \alpha_S x_{10} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma = 3)$. Thus, treatment increased the mean outcome by one unit. A binary outcome was generated using a previously described data-generating process for binary outcomes that induced data in which treatment caused a specified absolute risk reduction [16]. We simulated the binary outcome so that treatment caused a 0.02 reduction in the probability of the occurrence of outcome. Furthermore, the marginal probability of the occurrence of the event if all subjects were untreated was 0.10.

We generated a time-to-event outcome for each subject using a data-generating process for time-to-event outcomes based on the one described by Bender *et al.* [17]. For each subject, the linear predictor was defined 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: $U \sim (0, 1)$. A survival or event time was generated for each subjects as follows: $\left(\frac{-\log(u)}{\lambda e^{\text{LP}}}\right)^{1/\eta}$. 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 a previously described data-generating process for generating data with a specified marginal odds ratio or risk difference [16, 18]. 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) for those subjects who were ultimately assigned to receive the treatment (because matching allows one to estimate the average treatment effect in the treated—the ATT), we regressed the survival outcome on an indicator variable denoting treatment status. The coefficient for the treatment 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 outcomes model. A bisection approach was then employed to determine the value of β_{treat} that resulted in the desired marginal hazard ratio. We thus determined the value of β_{treat} that induced a desired marginal hazard ratio in the treated population. The true marginal hazard ratio was fixed at 0.8 for the current simulations.

We allowed the following factor to vary in our Monte Carlo simulations: the percentage of subjects that were treated (5%, 10%, 20%, and 25%). In each of the four scenarios, we simulated 1000 datasets, each consisting of 5000 subjects.

R code for generating the simulated datasets is provided in the Appendix.

3.2. Analyses in simulated datasets

Within each of the 1000 simulated datasets in each of the four scenarios, we estimated the propensity score using a logistic regression model to regress treatment status on the seven baseline covariates that affected the outcome. This approach to variable selection for the propensity-score model was selected, as it has been shown to result in better estimation compared with selecting only those variables that affect treatment selection [19]. We then used each of the three propensity-score matching algorithms described in Section 2.1 to form a propensity-score-matched sample. For a given matching method, the effect of treatment on the continuous, binary, and survival outcomes was determined using the methods described in Section 2.2 in each of the 1000 matched samples. The empirical sampling variability of the estimated treatment effects (difference in means, risk difference, and log-hazard ratio) was estimated as the standard deviation of the estimated treatment effects across the 1000 simulated datasets.

Within each of the 1000 matched samples, we used two different parametric methods to estimate the standard error of the estimated treatment effect. First, we used methods that accounted for the matched nature of the propensity-score-matched sample. These methods are described earlier in Section 2.2. Second, we used methods that did not account for the matched nature of the propensity-score-matched sample. These methods assumed that the treated and untreated subjects in the matched sample formed two independent samples. Thus, we used conventional methods for estimating the standard error of a difference in means and differences in proportions between two independent samples [20]. When outcomes were time-to-event in nature, a conventional Cox proportional hazards regression model, with model-based standard errors, was used to estimate the estimated log-hazard ratio and its standard error. The mean estimated standard error was then determined across the 1000 simulated datasets.

The simple bootstrap was then used to estimate the standard error of each estimated measure of effect. This was carried out as follows: (i) from the original matched sample in a given iteration of the simulation process, 1000 bootstrap samples of the matched pairs were drawn; (ii) the treatment effect was estimated in each of these 1000 bootstrap samples; and (iii) the standard error of the estimated treatment effect was estimated as the standard deviation of the estimated treatment effects across the 1000 bootstrap samples. Thus, in each of the 1000 iterations of the Monte Carlo simulation for a given scenario, we had a bootstrap estimate of the standard error of the estimated treatment effect. We then computed the mean bootstrap estimate of the standard error across the 1000 iterations of the simulations. Finally, the mean bootstrap estimate of the standard error across the 1000 iterations of the simulations was compared with the standard deviation of the estimated treatment effects across the 1000 simulated datasets.

Similarly, the complex bootstrap was used to estimate the sampling variability of each estimated measure of effect. This was carried out as follows: (i) from the original unmatched sample, 1000 bootstrap samples were drawn; (ii) the propensity score was estimated in each of these 1000 bootstrap samples using logistic regression; (iii) propensity-score matching was conducted in each of these 1000 bootstrap samples; (iv) the effect of treatment on each outcome was estimated in each of the 1000 propensity-score-matched samples; and (v) the standard error of the estimated treatment effect was estimated as the standard deviation of the estimated treatment effects across the 1000 bootstrap samples. Thus, in each of the 1000 iterations of the Monte Carlo simulation for a given scenario, we had a bootstrap estimate of the standard error of the estimated treatment effect. The mean bootstrap estimate of the standard error across the 1000 iterations was compared with the standard deviation of the estimated effects across the 1000 simulated datasets.

We estimated 95% confidence intervals for the estimated treatment effect in each of the 1000 datasets. When not using bootstrapping, standard normal-theory methods were used to estimate 95% confidence intervals based on the parametric variance estimates (naïve parametric variance estimate in the matched sample and the parametric method in the matched sample that accounted for the matched design). Three different methods were used to compute 95% confidence intervals when using the simple bootstrap: a standard normal-theory approach using the bootstrap estimate of the standard error of the estimated treatment effect, a nonparametric percentile-based estimate, and the BC_a estimate. Only the first two were used when using the complex bootstrap. We did not examine the use of the BC_a confidence intervals with the complex bootstrap because of the computational intensity of this approach (calculating the acceleration parameter a involves a leave-one-out jackknife-type procedure that does not lend itself well to simulations of matching within bootstrap samples). Matching tends to be more computationally intensive than drawing a random sample. Within a given iteration of the simulations, the simple bootstrap required that matching be carried out only once, while the complex bootstrap required that it be performed 1000 times. We then determined the empirical coverage rates of the estimated 95% confidence intervals by determin-

ing the proportion of the 1000 simulated datasets in which the estimated 95% confidence interval covered the true treatment effect used in the data-generating process.

The aforementioned Monte Carlo simulations used matching on the *estimated* propensity score: the propensity score was estimated in a given simulated dataset, and subjects were matched on this estimated propensity score. Both Rosenbaum [21] and Abadie and Imbens [22] suggest that matching on the estimated propensity score is more efficient than matching on the true propensity score. This appears to be because, while matching on the true propensity score adjusts for systematic differences in baseline characteristics between treatment groups, matching on the estimated propensity score adjusts for both systematic and random differences between treatment groups. To examine this effect, we repeated the aforementioned simulations using matching on the true propensity score (i.e., the propensity score that was used to simulate treatment status in the data-generating process). For this restricted set of simulations, we used B = 200 bootstrap samples and only examined standard normal-theory-based methods for estimating confidence intervals.

4. Monte Carlo simulations: results

4.1. Matching on the estimated propensity score

When the proportion of subjects who were treated was 0.05, 0.10, 0.20, and 0.25, then the average number of matched pairs formed using caliper matching across the 1000 simulated samples for each scenario was 248.8, 494.9, 967.0, and 1,175.8, respectively. Thus, on average, approximately 99.5%, 99.2%, 96.6%, and 94.0% of treated subjects were matched to an untreated subject. Thus, caliper matching should have minimal bias because of incomplete matching [5]. The two other matching methods, NNM and optimal matching, would result in all treated subjects being matched to an untreated subject.

The mean estimated standard errors and the empirical estimates of the standard deviation of the estimated treatment effects are reported in Figure 1. There is one panel for each of the three types of outcomes (continuous, binary, and time-to-event). Each panel consists of a series of dotcharts, with one row for

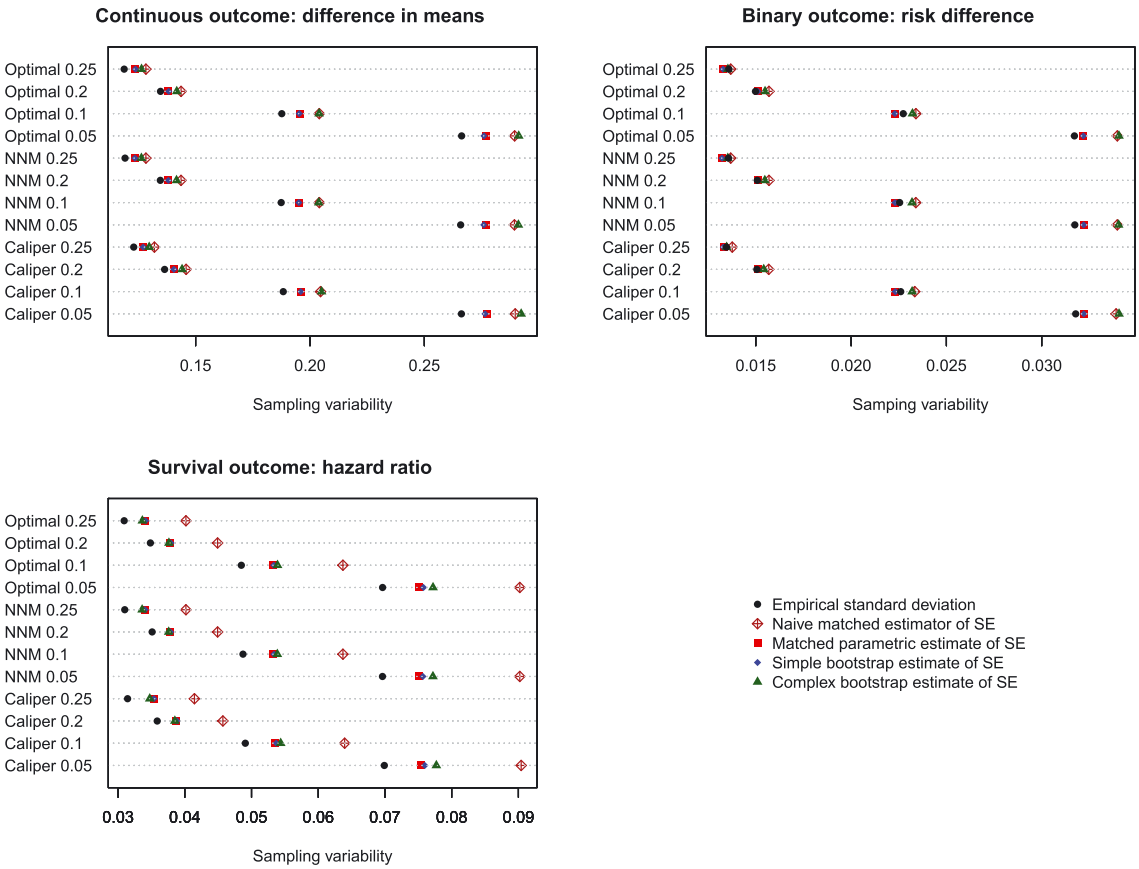

Figure 1. Empirical versus estimated standard errors (estimated propensity score).

each of the 12 combinations of matching method and prevalence of treatment (3 matching algorithms × 4 treatment prevalence). When outcomes were continuous, several observations merit comment. First, all four methods of estimating the standard error of the difference in means resulted in estimates that overestimated the standard deviation of the sampling distribution of the difference in means. Across the 12 combinations of prevalence of treatment (5%, 10%, 20%, and 25%) and the three matching methods, the mean ratio of the estimated standard error to the empirical standard error was 1.08 for the naïve matched estimator, 1.04 for the matched estimator that accounted for the matched nature of the sample, 1.04 for the simple bootstrap estimator of the standard error, and 1.07 for the complex bootstrap estimator of the standard error. Second, the naïve parametric estimator tended to result in estimates that resulted in the greatest overestimation of the variability of the sampling distribution. Third, the matched parametric estimator and the naïve bootstrap estimator tended to results in estimates that most closely reflected the empirical sampling variability of the estimated difference in means. In most settings, these two methods resulted in very similar estimates of standard error. Fourth, the complex bootstrap tended to have inferior performance compared with the simple bootstrap method. Fifth, results for the three different matching algorithms tended to be similar to one another. Sixth, differences between the methods for estimating the variability of the sampling distribution of the difference in means tended to diminish as the prevalence of treatment increased. Seventh, differences between the standard deviation of the empirical sampling distribution of the estimated difference in means and the mean estimated standard error for the four different methods tended to diminish as the prevalence of treatment increased. Thus, the estimates of standard error were most accurate when a higher proportion of subjects were treated. Similar results were observed when outcomes were binary, and the risk difference was used as the measure of treatment effect. When outcomes were binary, across the 12 combinations of prevalence of treatment and the three matching methods, the mean ratio of the estimated standard error to the empirical standard error was 1.04 for the naïve matched estimator, 1.00 for the matched estimator that accounted for the matched nature of the sample, 1.00 for the simple bootstrap estimator of the standard error, and 1.03 for the complex bootstrap estimator of the standard error. With time-to-event outcomes, similar results were observed with one primary exception. With time-to-event outcomes, the estimate of sampling variability obtained from the naïve parametric estimator was substantially larger than that obtained using the other three estimates.

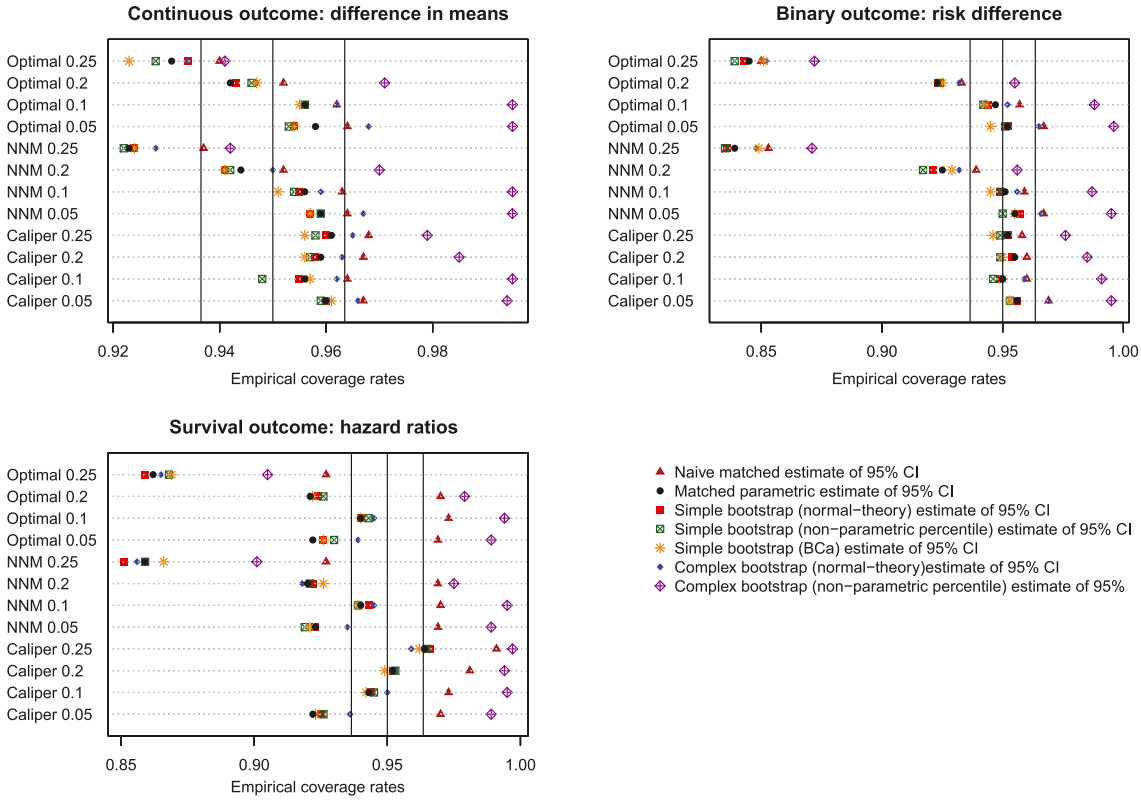

Figure 2. Empirical coverage rates of 95% confidence intervals (estimated propensity score).

When outcomes were time-to-event in nature, across the 12 combinations of prevalence of treatment and the three matching methods, the mean ratio of the estimated standard error to the empirical standard error was 1.30 for the naïve matched estimator, 1.09 for the matched estimator that accounted for the matched nature of the sample, 1.09 for the simple bootstrap estimator of the standard error, and 1.10 for the complex bootstrap estimator of the standard error.

The empirical coverage rates of estimated 95% confidence intervals obtained using the seven different methods (naïve parametric estimate, matched parametric estimate, simple bootstrap (normal-theory method), simple bootstrap (percentile method), simple bootstrap (BC_a), complex bootstrap (normal-theory method), and complex bootstrap (percentile method)) are reported in Figure 2. As above, there is a separate panel for each of the three different types of outcome (continuous, binary, and survival). Because of our use of 1000 simulated datasets for each scenario, an empirical coverage rate less than 0.9365 or greater than 0.9635 is statistically significantly different than the advertised rate of 0.95, based on a standard normal-theory test. On each panel, we have superimposed vertical lines denoting empirical coverage rates of 0.9365, 0.95, and 0.9635.

Several observations merit being highlighted. First, empirical coverage rates tended to be closer to the advertised rate of 95% for estimating means and risk differences compared with when estimating hazard ratios. Second, the simple bootstrap tended to result in better empirical coverage rates compared with the complex bootstrap. Third, coverage rates tended to be closer to the advertised rate when using caliper matching, compared with when NNM or optimal matching were used. Fourth, the naïve matched estimator tended to have worse performance compared with the matched parametric estimator and the simple bootstrap estimator. Fifth, the percentile-based confidence intervals did not perform well when used with the complex bootstrap.

4.2. Matching on the true propensity score

The corresponding results when we matched on the true propensity score are reported in Figures 3 and 4. In comparing Figures 1 and 3, one notes that there were minor differences in results when matching on the true propensity score compared with matching on the estimated propensity score. Several observa-

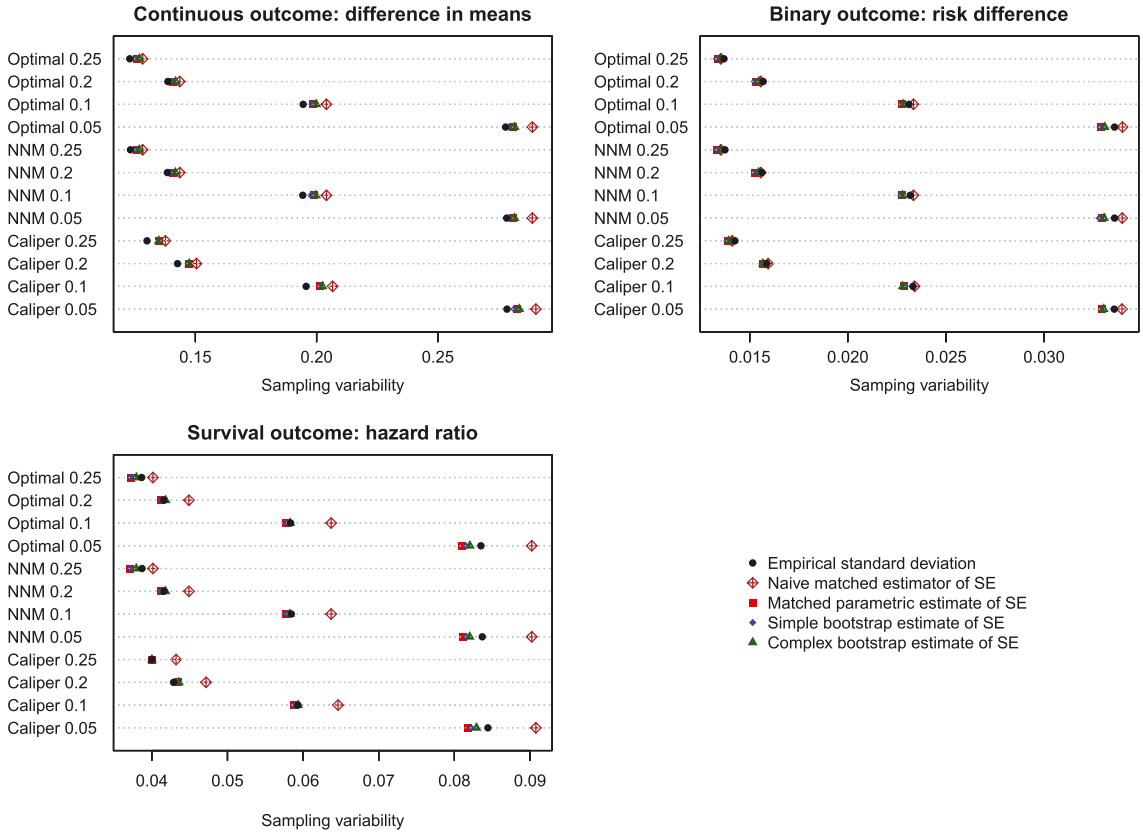

Figure 3. Empirical versus estimated standard errors (true propensity score).

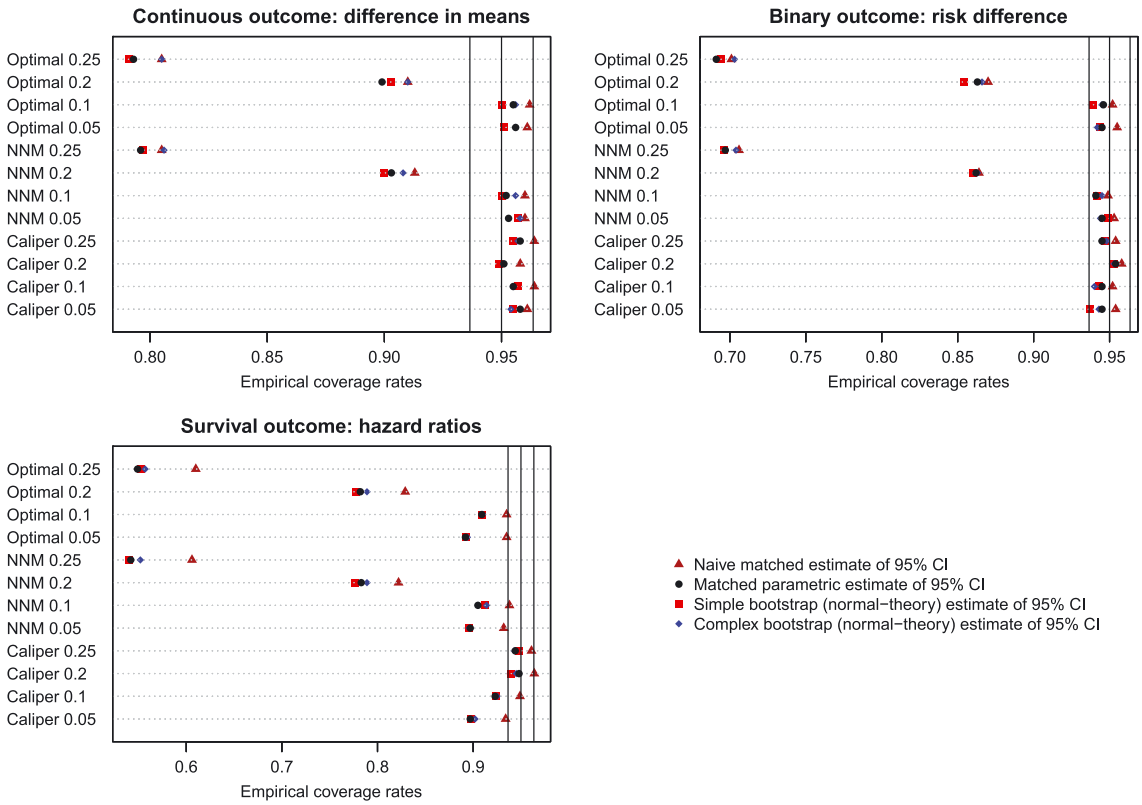

Figure 4. Empirical coverage rates of 95% confidence intervals (true propensity score).

tions merit being highlighted. First, the naïve matched estimator of the standard error tended to result in the greatest overestimate of the standard deviation of the empirical sampling distribution of the estimated effect compared with the three other competing methods. Second, the other three variance estimation methods (the matched parametric estimator and the two bootstrap methods) tended to result in estimates of the sampling variability that were almost indistinguishable from one another. Third, in most scenarios, these three variance estimation methods (the matched parametric estimator and the two bootstrap methods) tended to more closely approximate the sampling distribution compared with what was observed when matching on the estimated propensity score was used.

The empirical standard errors were smaller when the estimated propensity score was used compared with when the true propensity score was used. When estimating a difference in means, the ratio of the empirical standard error when using the estimated propensity score to the empirical standard error when using the true propensity score ranged between 0.94 and 0.97 across the different scenarios and methods. The comparable ranges were 0.94–0.94 for estimating a risk difference and 0.78–0.84 when estimating a marginal hazard ratio. This observations is in accordance with the suggestion of Rosenbaum [21] and Abadie and Imbens [22] that matching on the estimated propensity score is more efficient than matching on the true propensity score.

In examining Figure 4, one observes that when outcomes were continuous or binary, the use of caliper matching resulted in confidence intervals that had approximately correct coverage rates, regardless of the variance estimation method used, when matching on the true propensity score, compared with matching on the estimated propensity score. Coverage rates tended to be suboptimal when using optimal or NNM matching, and the prevalence of exposure was either 0.2 or 0.25.

5. Discussion

We conducted an extensive series of Monte Carlo simulations to examine the performance of bootstrap methods to estimate the variability of estimated treatment effects when using matching on the propensity score *without* replacement. We briefly summarize our findings and place them in the context of the existing literature.

Our primary focus was on examining the performance of two bootstrap methods for estimating the sampling variability of estimated treatment effects when matching *without* replacement on the estimated propensity score. We found that the simple bootstrap (bootstrapping matched pairs from the original propensity-score-matched sample) tended to result in variance estimates very similar to those obtained from a parametric variance estimate that accounted for the matched nature of the sample. A naïve variance estimate in the matched sample (i.e., one that disregarded the matched nature of the sample) tended to result in the greatest overestimation of the empirical sampling variability. In most cases, the variance estimation methods tended to overestimate the empirical sampling variability. As a secondary objective, we examined whether these observations changed when matching on the *true* propensity score, rather than on the *estimated* propensity score. When matching on the true propensity score, we observed that the parametric matched estimator and the two bootstrap estimators tended to result in very similar estimates of sampling variability. Furthermore, in many instances, these estimates were very close to the observed variability of the empirical distribution of estimated treatment effects. As above, when matching on the true propensity score, the use of the naïve variance estimator tended to result in the greatest overestimate of the sampling variability of the estimated treatment effect.

As noted in the Introduction, both Rosenbaum and Abadie and Imbens noted that matching on the estimated propensity score was more efficient than matching on the true propensity score [21, 22]. This is because in matching on the former score, one is removing both systematic and random differences in baseline characteristics between the two groups, while in matching on the latter score, one is only removing systematic differences between the two groups. This property of the estimated propensity score is likely the reason for the aberrant behavior of the complex bootstrap in our first set of simulations, in which we found that it had inferior performance to that of the simple bootstrap. Unlike in our simulations, the analyst in applied settings does not have the luxury of deciding whether to use the estimated or the true propensity score.

In a review article on nonparametric estimation of average treatment effects, Imbens briefly discussed the use of the bootstrap [23]. He suggested that if one is interested in the average treatment effect for the sample (as opposed to for the larger target population), then bootstrapping is inappropriate (p. 5). Accordingly, the use of the bootstrap should be restricted to contexts in which one is interested in making inferences about the effect of treatment in the larger population from which the sample was drawn. We would argue that this, in general, would be the case for many medical studies of treatment safety and efficacy and of epidemiological studies of the risks of exposures. Furthermore, he suggests that 'bootstrapping may be more complicated for matching estimators, as the process introduces discreteness in the distribution that will lead to ties in the matching algorithm' (p. 21). This issue was explored in greater depth in a subsequent paper by Abadie and Imbens [24]. They examined the validity of the standard bootstrap for NNM estimators *with replacement* and a fixed number of neighbors. They found that the standard bootstrap estimator was not valid for matching estimators in this context. It appears that one of the causes of the bias in the bootstrap estimator is that whenever a treated unit and the control unit to which the treated unit was originally matched both appear in the bootstrap sample, the treated unit is matched to the same control unit. This would not necessarily occur in our implementation of matching, because we only considered matching *without* replacement. Because of our use of matching *without* replacement, the findings of Abadie and Imbens do not necessarily hold. To the best of our knowledge, the current study represents the most comprehensive examination to date of the performance of bootstrapping for estimating variances in propensity-score-matched samples that were constructed using matching *without* replacement. Our findings suggest that the use of bootstrap-based methods is appropriate when using propensity-score matching *without* replacement. Given that, in the biomedical literature, matching *without* replacement is used substantially more frequently than matching *with* replacement, our results will be of interest to biostatisticians and applied health researchers. Our results help justify an additional methodological tool for use in the application of biostatistical methods. In the settings that we considered, we found that the use of simple bootstrap did not offer substantial advantages over the use of parametric estimators that accounted for the matched nature of the sample. Thus, in many settings, the use of the bootstrap may not be necessary. However, in more complex settings in which a parametric estimate has not been developed, the use of the bootstrap may offer substantial advantage. An example is the recently-proposed method of double propensity score adjustment [25].

In the current study, we only considered methods that used matching without replacement and did not consider methods that used matching with replacement. Hill and Reiter described methods to estimate the standard error of the estimated treatment effect when using matching with replacement [26]. However, their described estimator is only applicable when estimating linear treatment effects for continuous outcomes. Methods for estimating standard errors when matching with replacement and when outcomes are binary or time-to-event in nature have not been described. Based on the conclusions of Abadie and Imbens [24], bootstrap methods can result in biased estimation of the standard error of estimated effects when using matching with replacement. Accordingly, the use of the bootstrap is unlikely to circumvent difficulties due to the absence of formulas for the standard error of estimated risk differences, relative risks, or hazard ratios when matching with replacement is used. However, in our experience, matching with replacement is rarely used in the biomedical literature, where the use of matching without replacement appears to predominate.

We examined three different matching algorithms: optimal matching, NNM, and caliper matching. Estimation of the standard error of the estimated treatment effect was very similar between the first two methods. Given earlier research, this finding was not surprising. Gu and Rosenbaum found that optimal matching and NNM resulted in comparable balance in measured baseline covariates between treatment groups [27]. Similarly, a recent study found that these two methods resulted in estimates with similar bias, variance, and mean squared error [28]. Despite these prior studies demonstrating the similar performance of optimal matching and NNM, optimal matching was included in the current study for the sake of completeness.

Our findings, combined with the theoretical derivations of Abadie and Imbens, have practical implications for users of statistical software programs for implementing propensity-score matching. First, in deriving the variance of the matched estimator, some statistical software programs may assume that outcomes are independent across units. In our simulations, we found that the naïve matched estimator that assumed independent observations tended to result in the greatest bias in estimating the sampling variability of the estimated treatment effect. Similar findings have been observed in prior studies [7–10]. Second, some statistical software packages may use matching *with* replacement as the default option. Based on the results of Abadie and Imbens, the use of the bootstrap in this context may result in biased estimation of the sampling variability of the matched estimator because of the use of matching with replacement. Users of any user-written software packages (indeed of all statistical software) need to be aware of the specific implementation that is used and the advantages and limitations of that implementation. Different matching algorithms and different approaches to variance estimation cannot be considered interchangeable.

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 number and distribution of baseline covariates and about their relationship with both treatment selection and with the outcome. Given our use of propensity-score matching, analytic determination of the performance of bootstrap methods for variance estimation would be very difficult. We note that several prior studies examining the performance of propensity-score methods for estimating treatment effects have employed Monte Carlo simulations [29–34]. A second limitation is our focus on the use of pair matching, in which pairs of treated and untreated subjects were formed. We did not consider alternative matching algorithms such as full matching [35]. Our focus on pair matching is justified because it is the method that is used most frequently in the biomedical literature [2–4]. A further strength of our study was that we considered two different greedy algorithms for matching as well as an optimal matching algorithm. Thus, our consideration of matching algorithms was more comprehensive than those in the majority of studies that used simulations to examine different aspects of propensity-score matching.

Our simulations found that the simple bootstrap resulted in estimates of the standard error that were often very similar to the parametric-based estimators of the standard error that accounted for the matched nature of the sample. Given these findings, we suggest that parametric-based estimates of the standard error that account for the matched nature of the sample be used when such estimators exist. The fact that the simple bootstrap performed similar to parametric-based estimators that accounted for the matched nature of the sample suggests that the simple bootstrap may be useful for more complex outcomes or more complex estimates of treatment effect than are considered in this paper, for which parametric-based estimators of the standard error that account for matching have not been developed.

Appendix

R statistical software code for generating the simulated datasets

```
set.seed(iter)
# Random number seed is set within each iteration so that the results
# are reproducible.

 N <- 5000
 # Size of simulated dataset.

 beta.0.treat <- scan("beta.0.treat.out")
 # The intercept in the treatment-selection model. This will determine the
 # prevalence of treatment in the simulated dataset. The appropriate value
 # of the intercept can be found using a grid search or a bisection approach.

 beta.0.outcome <- scan("beta.0.outcome.out")
 # The intercept in the binary-outcome generating model.
 # Its value will determine the incidence of the outcome.
 # The appropriate value of the intercept can be found using a bisection
 # approach.

 beta.effect <- scan("beta.effect.out")
 # The log-odds ratio for the effect of treatment on the outcome that will
 # induce the desired marginal risk difference. It can be found using a
 # bisection approach.

 beta.hr <- scan("beta.hr.out")
 # The conditional log-hazard ratio for the effect of treatment that will
 # induce the desired marginal hazard ratio. It can be found using a
 # bisection approach.

 x.1 <- rnorm(N,0,1)
 x.2 <- rnorm(N,0,1)
 x.3 <- rnorm(N,0,1)
 x.4 <- rnorm(N,0,1)
 x.5 <- rnorm(N,0,1)
 x.6 <- rnorm(N,0,1)
 x.7 <- rnorm(N,0,1)
 x.8 <- rnorm(N,0,1)
 x.9 <- rnorm(N,0,1)
 x.10 <- rnorm(N,0,1)

 beta.low <- log(1.25)
 beta.med <- log(1.5)
 beta.high <- log(1.75)
 beta.v.high <- log(2)

# Generate treatment status for each subject.
logit.treat <- beta.0.treat + beta.low*x.1 + beta.med*x.2 +
  beta.high*x.3 + beta.low*x.4 + beta.med*x.5 + beta.high*x.6 +
  beta.v.high*x.7

p.treat <- exp(logit.treat)/(1 + exp(logit.treat))

treat <- rbinom(N,1,p.treat)
```

```
# Generate continuous outcome for each subject.
# Generate a continuous outcome with a treatment effect of 1.

y <- 1*treat + beta.low*x.4 + beta.med*x.5 + beta.high*x.6 +
 beta.v.high*x.7 + beta.low*x.8 + beta.med*x.9 +
 beta.high*x.10 + rnorm(N,0,3)

# Generate a binary outcome for each subject.
# Generate a binary outcome with a treatment effect of a risk difference of
# 0.02 (10% of subjects will have outcome if untreated).

logit.outcome <- beta.0.outcome + beta.effect*treat +
 beta.low*x.4 + beta.med*x.5 + beta.high*x.6 +
 beta.v.high*x.7 + beta.low*x.8 + beta.med*x.9 +
 beta.high*x.10
p.outcome <- exp(logit.outcome)/(1 + exp(logit.outcome))

outcome <- rbinom(N,1,p.outcome)

# Generate a time-to-event for each subject. True hazard ratio is 0.8.

linpred <- beta.hr*treat +
 beta.low*x.4 + beta.med*x.5 + beta.high*x.6 +
 beta.v.high*x.7 + beta.low*x.8 + beta.med*x.9 +
 beta.high*x.10

lambda <- 0.00002
nu <- 2

ranu <- runif(N,min=0,max=1)

surv.time <- ( -(log(ranu))/(lambda*exp(linpred)) )^(1/nu)
surv.status <- rep(1,N)
```

Acknowledgements

This study was supported by the Institute for Clinical Evaluative Sciences (ICES), which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by 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. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
2. 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.
3. 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.
4. 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.
5. 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.
6. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological Methods* 2008; **13**(4):279–313.
7. 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.

8. 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).
9. 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.
10. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stastisics in Medicine* 2013; **32**(16):2837–2849.
11. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, NY, 1993.
12. 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.
13. Rosenbaum PR. *Observational Studies*. Springer-Verlag: New York, NY, 2002.
14. Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine* 2004; **23**(1):65–75.
15. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *Journal of the American Statistical Association* 1989; **84**:1074–1078.
16. 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.
17. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723.
18. 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.
19. 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.
20. Rosner B. *Fundamentals of Biostatistics*. Duxbury Press: Belmont, CA, 1995.
21. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**:387–394.
22. Abadie A, Imbens GW. Matching on the estimated propensity score. *National Bureau of Economic Research* 2009; **NBER Working Paper Series**(15301).
23. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* 2004; **86**:4–29.
24. Abadie A, Imbens GW. Notes and comments on the failure of the bootstrap for matching estimators. *Econometrica* 2008; **76**(6):1537–1557.
25. Austin PC. Double propensity-score adjustment: A solution to bias due to incomplete matching. *Statistical Methods in Medical Research* In-press. DOI: 10.1177/0962280214543508.
26. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine* 2006; **25**(13):2230–2256.
27. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 1993; **2**:405–420.
28. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Statisics in Medicine* 2013; **33**(6): 1057–1069.
29. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**(3):337–346.
30. 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.
31. 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.
32. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine* 2007; **26**(16):3078–3094.
33. Austin PC. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology* 2008; **61**(6):537–545.
34. 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.
35. Hansen BB. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 2004; **99**(467):609–618.

