Statistics
in Medicine

Research Article

Received: 9 March 2016, Accepted: 1 August 2016 Published online 22 August 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7084

Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis

Peter C. Austin[a,b,c*†]

Propensity score methods are used to reduce the effects of observed confounding when using observational data to estimate the effects of treatments or exposures. A popular method of using the propensity score is inverse probability of treatment weighting (IPTW). When using this method, a weight is calculated for each subject that is equal to the inverse of the probability of receiving the treatment that was actually received. These weights are then incorporated into the analyses to minimize the effects of observed confounding. Previous research has found that these methods result in unbiased estimation when estimating the effect of treatment on survival outcomes. However, conventional methods of variance estimation were shown to result in biased estimates of standard error. In this study, we conducted an extensive set of Monte Carlo simulations to examine different methods of variance estimation when using a weighted Cox proportional hazards model to estimate the effect of treatment. We considered three variance estimation methods: (i) a naïve model-based variance estimator; (ii) a robust sandwich-type variance estimator; and (iii) a bootstrap variance estimator. We considered estimation of both the average treatment effect and the average treatment effect in the treated. We found that the use of a bootstrap estimator resulted in approximately correct estimates of standard errors and confidence intervals with the correct coverage rates. The other estimators resulted in biased estimates of standard errors and confidence intervals with incorrect coverage rates. Our simulations were informed by a case study examining the effect of statin prescribing on mortality. © 2016 The Authors. *Statistics in Medicine* published by John Wiley & Sons Ltd.

Keywords: propensity score; survival analysis; inverse probability of treatment weighting (IPTW); Monte Carlo simulations; variance estimation; observational study

1. Introduction

Observational studies are increasingly being used to estimate the effects of treatments, interventions and exposures on outcomes. Confounding is an important issue in such studies as treated subjects often differ systematically from control subjects in terms of baseline covariates that are prognostically important. Statistical methods must be used to remove or minimize the effect of confounding due to measured covariates so that valid inferences on treatment effects can be drawn from observational studies.

An important class of bias-reduction methods are those that are based on the propensity score [1]. The propensity score is the probability of treatment assignment conditional on measured baseline covariates. These methods are increasingly being used to reduce or minimize the confounding that occurs in observational studies. 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 [1–3]. These methods are frequently used in the biomedical literature [4,5].

Survival or time-to-event outcomes occur frequently in the medical and epidemiological literature [6]. Recent studies have examined the performance of different propensity score methods for estimat-

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

ing the effect of treatment on survival outcomes [7,8]. In one study, it was shown that both matching on the propensity score and IPTW using the propensity score resulted in unbiased estimation of marginal hazard ratios [8]. However, IPTW using the propensity score with the conventional variance estimator resulted in biased estimation of standard errors and confidence intervals whose empirical coverage rates differed from the advertised rates. Variance estimation when matching on the propensity score [7,9–12] or when using IPTW with linear treatment effects [13] has received a moderate degree of attention. However, little attention has been focused on variance estimation when using IPTW with survival outcomes.

The objective of the current study was to examine variance estimation when using IPTW using the propensity score to estimate the effect of treatment on survival or time-to-event outcomes when using a Cox proportional hazards model. The paper is structured as follows: In Section 2, we briefly describe different methods for variance estimation when using IPTW using the propensity score with survival outcomes. In Section 3, we introduce a brief case study illustrating the application of these methods. In Section 4, we describe an extensive series of Monte Carlo simulations that were performed to compare the performance of different variance estimators. In Section 5, we report the results of these simulations. Finally, in Section 6, we summarize our findings and place them in the context of the existing literature.

2. Using propensity score weighting to estimate the effect of treatment on survival outcomes

We use the following notation throughout this section. Let Z denote an indicator variable denoting treatment status ($Z=1$ for active treatment of interest vs. $Z=0$ for the control treatment), while e denotes the propensity score. The propensity score is typically estimated using a logistic regression model in which the binary indicator variable denoting treatment status is regressed on observed baseline covariates. Alternatively, methods from the machine learning literature, such as random forests or generalized boosting methods can be used [14–16]. Variable selection for the propensity score model has been considered elsewhere [17].

The inverse probability of treatment weights (IPTWs) are defined as $w_{\text{ate}} = \frac{Z}{e} + \frac{1-Z}{1-e}$ [13]. Thus, each subject is weighted by the reciprocal of the probability of receiving the treatment that the subject actually received. Weighting the sample using these weights results in a synthetic sample in which observed baseline covariates are not confounded with treatment assignment. The use of these weights allows one to estimate the average treatment effect (ATE). We refer to these weights as the conventional IPTW-ATE weights. An alternative to the conventional IPTW-ATE weights are the stabilized IPTW-ATE weights: $w_{\text{ate,stab}} = \Pr(Z=1)\frac{Z}{e} + \Pr(Z=0)\frac{1-Z}{1-e}$ [18,19]. In defining the stabilized weights, the reciprocal of the probability of receiving a given treatment is multiplied by the marginal probability of receiving the given treatment. Stabilized weights are intended to reduce variability due to instability in estimation that can be induced by subjects with very large weights. Using weights equal to $w_{\text{att}} = Z + \frac{e(1-Z)}{1-e}$ allows one to estimate the average treatment effect in the treated (ATT) [20]. We refer to these weights as IPTW-ATT weights. With these weights, treated subjects receive a weight of one, while the control subjects receive a weight of the odds of receiving the active treatment. Thus, the population of treated subjects serves as the reference population to which each of the treated and control populations are standardized.

To estimate the effect of treatment on the hazard of the occurrence of the outcome, one can use a weighted Cox regression model to regress survival on an indicator variable denoting treatment status [8,21]. Three possible methods can be used to estimate the variance of the estimated treatment effect. First, one could use the naïve model-based variance estimator from the maximum partial likelihood estimator for the Cox proportional hazards model. Second, one could use a robust sandwich-type variance estimator of the type proposed by Lin [22]. The use of this estimator with IPTW regression models was proposed by Joffe et al. [21]. A rationale for the use of a robust variance estimator is provided by Hernan et al., who note that the use of weights induces a within-subject correlation in outcomes as observations can have weights that are unequal to one another [23,24]. Xu et al. showed that the use of conventional IPTW-ATE weights tended to result in a doubling of the number of subjects in the analytic sample [25]. The use of a robust variance estimator accounts for the lack of independence in replications of subjects induced by weighting. The third approach to variance estimation is to use a bootstrap-based method to estimate the variability of the estimated treatment effect. Of these three variance estimation methods, the

use of a robust variance estimator appears to be the most frequent. However, the relative performance of the three different methods is not known.

Statistical software code in R and SAS for fitting the weighted Cox regression model with robust standard errors is provided in the appendix.

3. Case study

We provide an empirical example to illustrate the application of the methods described in the previous section.

3.1. Data and analyses

We used data from a previously published tutorial on the use of propensity score methods with survival data [26,27]. The sample consisted of 9107 subjects patients discharged from hospital with acute myocardial infarction (AMI) in Ontario, Canada. Data on patient characteristics were obtained by retrospective chart review by trained cardiovascular research nurses. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment Study [28].

The exposure of interest was whether the patient received a prescription for a statin lipid lowering agent at hospital discharge. Three thousand and forty-nine subjects (33.5%) received a statin prescription at hospital discharge. Patients were followed for up to eight years post-discharge. Patients who survived to eight years post-discharge had their survival times treated as censored observations.

For the purposes of this case study, we considered 10 baseline characteristics: age, systolic blood pressure at admission, respiratory rate at admission, creatinine, history of previous cardiovascular revascularization procedure (percutaneous coronary intervention or coronary artery bypass graft surgery), diabetes, current smoker, history of hyperlipidemia, history of dementia and history of a previous AMI. The first four are continuous covariates while the last six are binary.

Statin exposure was regressed on the set of 10 baseline covariates using a logistic regression model. All 10 covariates were statistically significantly associated with statin prescribing at hospital discharge ($P < 0.032$). The hazard of post-discharge mortality was regressed on the 10 covariates using a Cox proportional hazards regression model. All 10 covariates were significantly associated with the hazard of death ($P < 0.002$). The corresponding odds ratios (for exposure) and hazard ratios (for death) are reported in Table I.

The propensity score was estimated for each subject using the logistic regression model estimated above. The three sets of weights described in Section 2 were calculated (conventional IPTW-ATE weights, stabilized IPTW-ATE weights, IPTW-ATT weights). The effect of statin prescribing on the hazard of post-discharge mortality was estimated using a weighted Cox proportional hazards model. The three different variance estimators described above were used to construct 95% confidence intervals for the effect of treatment on the hazard of mortality.

When using bootstrapping, 200 bootstrap samples were drawn from the study sample [29] (Efron and Tibshirani suggest that 200 bootstrap samples are generally sufficient for estimating a standard error (page 52)). In each bootstrap sample, the propensity score was estimated by regressing treatment status

Table I. Odds ratios and hazard ratios for treatment selection and mortality.

Variable	Odds ratio for statin prescribing (95% CI)	Hazard ratio for death (95% CI)
Age (per year increase)	0.980 (0.976, 0.984)	1.075 (1.071, 1.078)
Systolic blood pressure (per mmHg increase)	1.002 (1.001, 1.004)	0.996 (0.995, 0.997)
Respiratory rate (per breath per minute)	0.985 (0.975, 0.995)	1.036 (1.031, 1.041)
Creatinine (per μmol/L)	0.999 (0.998, 1.000)	1.003 (1.003, 1.003)
Previous revascularization	1.323 (1.117, 1.568)	1.201 (1.074, 1.343)
Diabetes	0.849 (0.758, 0.951)	1.766 (1.647, 1.893)
Current smoker	0.875 (0.783, 0.977)	1.247 (1.149, 1.355)
Hyperlipidemia	5.307 (4.804, 5.863)	0.850 (0.785, 0.92)
Dementia	0.488 (0.331, 0.72)	1.727 (1.508, 1.978)
Previous AMI	1.196 (1.056, 1.354)	1.424 (1.321, 1.534)

on the 10 baseline covariates using a logistic regression model. The hazard of death was regressed on an indicator variable denoting treatment status using a weighted Cox model fit to the bootstrap sample. This was done for the three different sets of weights. The standard deviation of the estimated log-hazard ratios (i.e., the estimated regression coefficient for the treatment status indicator) across the 200 bootstrap samples was used as the bootstrap estimate of the standard error of the estimated regression coefficient obtained in the original analytic sample. Ninety-five percent confidence intervals were constructed by as $\hat{\beta} \pm 1.96 \times \mathrm{Se}(\hat{\beta})$, where $\hat{\beta}$ denotes the estimated effect in the original analytic sample, and $\mathrm{Se}(\hat{\beta})$ denotes the estimated standard error of the estimated treatment effect (obtained using either the naïve variance estimator, the robust sandwich-type variance estimator, or bootstrapping).

3.2. Results

The nine estimated hazard ratios with associated 95% confidence intervals (three different sets of weights × three methods for variance estimation) are reported in Table II. Several observations warrant comment. First, the estimates of the ATE are further from the null than are the estimates of the ATT (hazard ratios of 0.76 vs. 0.90). An interpretation of this result is that the effect of statin prescribing on mortality is stronger in the entire AMI population than it is in those AMI patients who ultimately received a statin at hospital discharge. Second, when estimating the ATE, variance estimates were substantially smaller when a naïve variance estimator was used compared to when a robust variance estimator or a bootstrap estimator was used. Third, differences between the robust variance estimator and the bootstrap estimator were negligible. Fourth, when estimating the ATE, the choice of weight (IPTW-ATE weight vs. the stabilized ATE weight) had essentially no impact upon the estimate of the point estimate of statin efficacy.

4. Monte Carlo simulations—methods

We used an extensive series of Monte Carlo simulations to examine the performance of different variance estimators when using IPTW with survival outcomes. We based our simulations on the empirical analyses conducted in the previous section so that our simulations would reflect the empirical data on which the case study was based.

We simulated data for a setting in which there were 10 baseline covariates ($X_1 - X_{10}$). The first four were continuous while the last six were binary (to reflect the setting of the case study). Using the data from the case study, we standardized the four continuous covariates so that they had mean zero and unit variance. We estimated the Pearson variance–covariance matrix of the 10 baseline covariates. We simulated 10 baseline covariates for each subject from a multivariate normal distribution with mean vector equal to zero and with variance–covariance matrix equal to that estimated above. Each of the last six simulated covariates was used to create a binary variable with a prevalence equal to that of one of the binary covariates in the case study. This was done by determining whether the simulated normally distributed variable lay above or below a specified quantile of the given normal distribution.

Table II. Estimated treatment effects, variance estimates and 95% confidence intervals from case study.

Estimation method	Estimated log-hazard ratio (standard error)	Estimated hazard ratio (95% CI)
	Estimates of the ATE	
IPTW weights – naïve variance estimate	−0.2729 (0.0248)	0.76 (0.73, 0.80)
IPTW weights – robust variance estimate	−0.2729 (0.0436)	0.76 (0.70, 0.83)
IPTW weights – bootstrap variance estimate	−0.2729 (0.0425)	0.76 (0.70, 0.83)
Stabilized weights – naïve variance estimate	−0.2722 (0.0378)	0.76 (0.71, 0.82)
Stabilized weights – robust variance estimate	−0.2722 (0.0435)	0.76 (0.70, 0.83)
Stabilized weights – bootstrap variance estimate	−0.2722 (0.0459)	0.76 (0.70, 0.83)
	Estimates of the ATT	
Naïve variance estimation	−0.1082 (0.0463)	0.90 (0.82, 0.98)
Robust variance estimation	−0.1082 (0.0461)	0.90 (0.82, 0.98)
Bootstrap variance estimation	−0.1082 (0.0488)	0.90 (0.82, 0.99)

As in the case study, we assumed that each of the 10 simulated baseline covariates influenced treatment assignment. For each subject, the probability of treatment selection was determined from the following logistic model: $\text{logit}(p_i) = \alpha_{0,\text{treat}} + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5 + \alpha_6 x_6 + \alpha_7 x_7 + \alpha_8 x_8 + \alpha_9 x_9 + \alpha_{10} x_{10}$. A bisection approach, as described in previous publications, was used to determine 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 [8]. The regression coefficients ($\alpha_1, \ldots, \alpha_{10}$) were set to the values estimated using the logistic regression model in the case study (with the modification that the continuous variables had been standardized to mean zero and unit variance). For each subject, treatment status was generated from a Bernoulli distribution with subject-specific parameter: $Z \sim \text{Be}(p_i)$.

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. [30]. For each subject, the linear predictor was defined as

$$\text{LP} = \beta_{\text{treat}} Z + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10}.$$

As in the case study, the 10 baseline covariates all affected the hazard of the outcome. The regression coefficients ($\beta_1, \ldots, \beta_{10}$) were set to the values estimated in the case study (with the modification that the continuous variables had been standardized to mean zero and unit variance). The value of the log-hazard ratio for the treatment indicator (β_{treat}) was allowed to vary in the Monte Carlo simulations. For each subject, we generated a random number from a standard uniform distribution: $u \sim \text{U}(0,1)$. An event time was generated for each subject 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, as was done in previous studies [8,31].

The use of this data-generating process results in a conditional treatment effect, with a conditional hazard ratio of $\exp(\beta_{\text{treat}})$. However, IPTW using the propensity score estimates the marginal or population-average effect. To determine the true marginal hazard ratio (this was necessary to permit estimation of empirical coverage rates of estimated confidence intervals), we generated a very large dataset consisting of 1 000 000 subjects using the methods described in the previous paragraphs with one important modification. For each subject, we simulated two outcomes: the two potential outcomes under the treatment and control conditions. Thus, we simulated an outcome for each subject under the assumption that they were treated. We then simulated an outcome for each subject under the assumption that they were under the control condition. Then in the dataset consisting of both outcomes for each subject, we regressed the hazard of the outcome on the indicator variable denoting treatment status. The estimated log-hazard ratio will serve as the true marginal ATE. We then repeated the above regression analysis, restricting the analysis sample to those subjects who were ultimately assigned to the treatment in the large simulated population. The estimated log-hazard ratio will serve as the true marginal ATT.

We allowed the following factors to vary in our Monte Carlo simulations: the percentage of subjects that were treated (5%, 10%, 20%, 30%, 40%, and 50%) and the true conditional hazard ratio (1, 2, 3, 4, and 5). We thus examined 30 scenarios (6 treatment prevalences × 5 conditional hazard ratios). In each scenario, we simulated 1000 datasets, each consisting of 1000 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 10 baseline covariates. In each of the 1000 simulated datasets for each scenario, we estimated the log-hazard ratio and its standard error using the methods described in Section 2. Thus, we used three different sets of weights: the conventional IPTW-ATE weights, the stabilized IPTW-ATE, and the IPTW-ATT weights. We used three different variance estimators: the naïve model-based estimates, the robust sandwich-type variance estimators, and a bootstrap estimate based on 200 bootstrap samples. Let $\hat{\theta}_i$ and $\hat{\gamma}_i$ denote the estimated log-hazard ratio and its estimated standard error, respectively, obtained from the ith simulated dataset using a given method. The estimated 95% confidence interval for the estimated log-hazard ratio was computed as $\hat{\theta}_i \pm 1.96 \times \hat{\gamma}_i$.

We examined the accuracy of variance estimation three different ways. First, we 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 quantities. We determined the mean standard error of the log-hazard ratio across the 1000 iterations: $\bar{\hat{\gamma}} = \frac{1}{1{,}000} \sum_{i=1}^{1{,}000} \hat{\gamma}_i$. We also determined the standard deviation of the estimated log-hazard ratios across the 1000 simulated datasets: $\text{sd}(\hat{\theta}_i)$. The first quantity estimates the mean standard error, while the second quantity is the empirical

estimate of the sampling variability of the log-hazard ratio. We then determined the ratio of these two quantities: $\frac{\bar{\hat{\gamma}}}{\mathrm{sd}(\hat{\theta}_i)}$. If the ratio equals one, then the estimated standard error of the log-hazard ratio is correctly estimating the sampling variability of the estimated log-hazard ratio. Second, we determined the proportion of estimated 95% confidence intervals that covered the true log-hazard ratio across the 1000 iterations of each scenario. Third, we determined the mean length of the estimated 95% confidence intervals across the 1000 iterations of each scenario.

5. Monte Carlo simulations—results

The results of the Monte Carlo simulations are reported in Figures 1 through 6. The first three figures summarize results for estimation of the ATE, while the last three figures summarize results for the estimation of the ATT. Each figure consists of six panels, with one panel for each of the true conditional hazard ratios used in the data-generating process.

5.1. Estimation of the ATE

The ratio of the mean estimated standard error to the empirical standard deviation of the sampling distribution of the log-hazard ratio is reported in Figure 1. In general, the bootstrap variance estimator resulted in a ratio that was approximately equal to one, indicating that the bootstrap estimator correctly

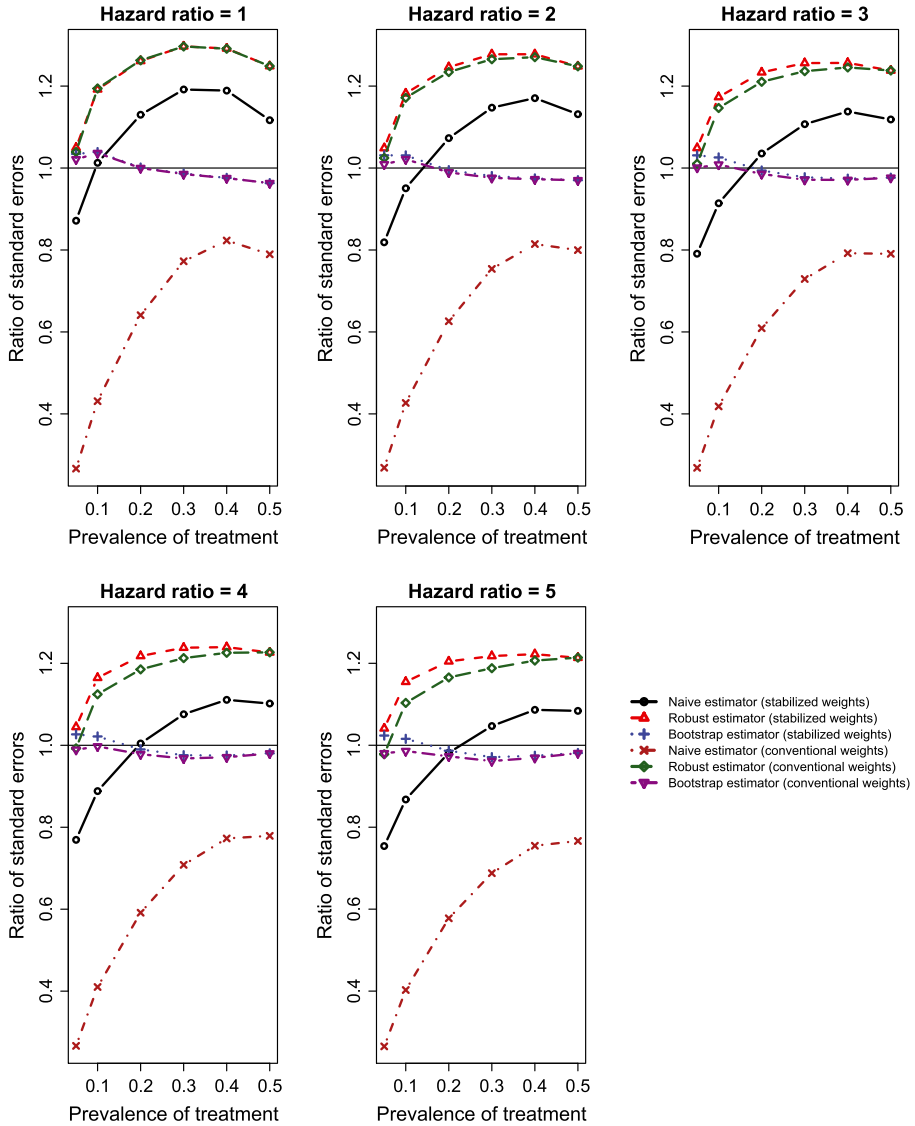

Figure 1. Ratio of mean estimated standard error to empirical standard deviation of sampling distribution: ATE.

approximated the standard deviation of the empirical sampling distribution of the log-hazard ratio. This was true regardless of whether the conventional IPTW-ATE or the stabilized IPTW-ATE weights were used. The naïve variance estimator with the conventional IPTW-ATE weights tended to substantially under-estimate the standard deviation of the sampling distribution of the log-hazard ratio. The remaining estimators tended to over-estimate the sampling variability, particularly when the prevalence of treatment was above 25%. It is important to note that, in general, the robust variance estimator over-estimated the sampling variability of the log-hazard ratio, regardless of which set of weights was used.

The empirical coverage rates of the estimated 95% confidence intervals are reported in Figure 2. Due to our use of 1000 iterations within each scenario, any empirical coverage rate that exceeded 0.9635 or that was less than 0.9365 would be statistically significantly different from the nominal rate of 0.95 based on a standard normal-theory test. Horizontal lines denoting these thresholds, along with the nominal rate of 0.95 have been superimposed on each panel. The use of the naïve variance estimator with the conventional IPTW-ATE weights resulted in estimated 95% confidence intervals whose empirical coverage rates were substantially lower than the nominal level. The two bootstrap estimators (using conventional IPTW-ATE weights or the stabilized IPTW-ATE weights) tended to result in estimates with approximately correct coverage rates. The remaining three methods (robust estimator with conventional IPTW-ATE weights, naïve estimator with stabilized IPTW-ATE weights, robust estimator with stabilized IPTW-ATE weights) frequently resulted in confidence intervals whose empirical coverage rates exceeded the nominal level.

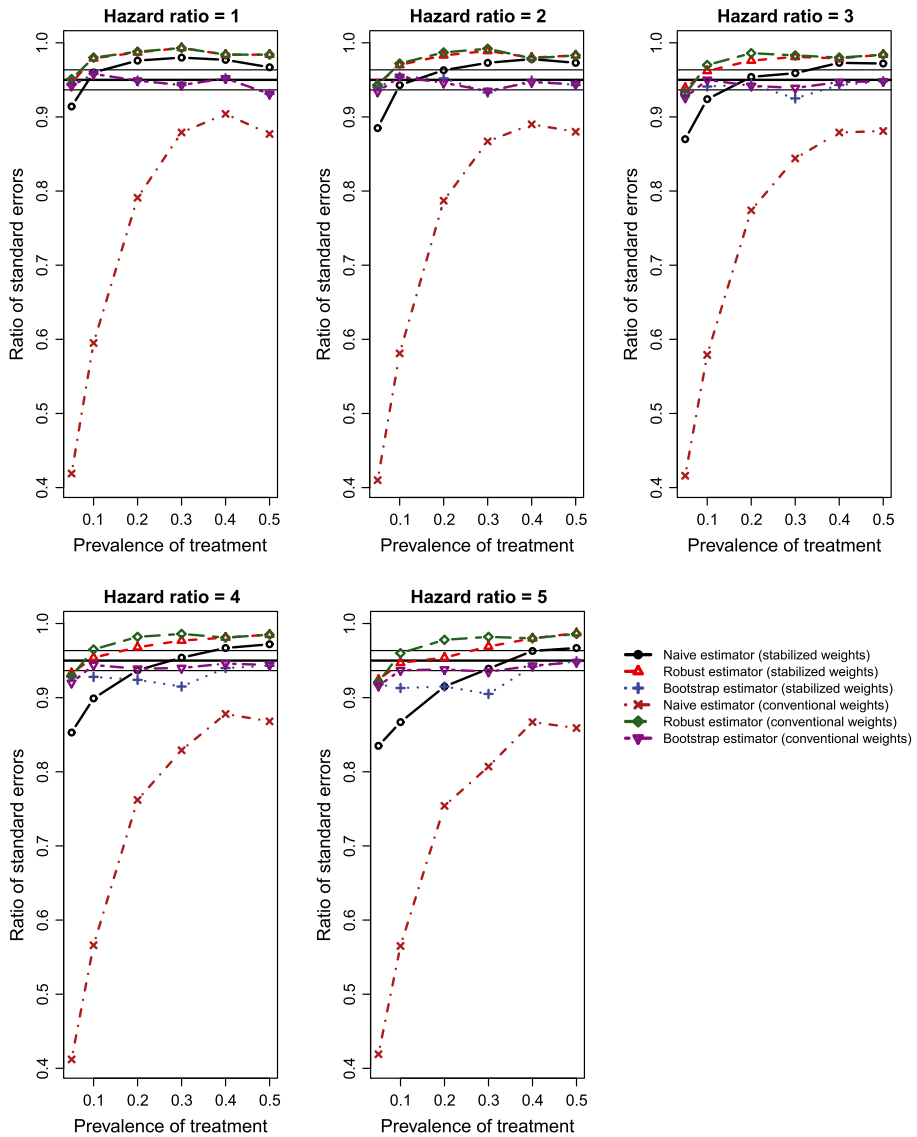

Figure 2. Empirical coverage rates of estimated 95% confidence intervals: ATE.

Because of the equivalence between confidence intervals covering a given value and hypothesis testing, the scenario in which the true hazard ratio was equal to one permits us to examine the type I error rate. The empirical type I error rate is equal to the proportion of estimated confidence that exclude the true value. In examining the upper left panel of Figure 2, one can see that the two bootstrap-based estimators tended to have approximately correct type I error rates, while the other four methods tended to have incorrect type I error rates. In particular, the use of conventional IPTW-ATE weights with a naïve variance estimator tended to result in substantially inflated type I error rates. The other three methods tended to result in type I error rates that were lower than the expected rate of 5%.

The mean length of estimated 95% confidence intervals are reported in Figure 3. As was to be expected, based on the results reported above, the use of naïve variance estimator with the conventional IPTW-ATE weights resulted in the narrowest confidence intervals. Among the remaining five estimators, the two bootstrap estimators (with the conventional IPTW-ATE weights or with the stabilized IPTW-ATE weights) tended to result in the narrowest 95% confidence intervals.

5.2. *Estimation of the ATT*

The degree to which the estimated standard error correctly approximated the standard deviation of the empirical sampling distribution of the log-hazard ratio is reported in Figure 4. The use of the bootstrap

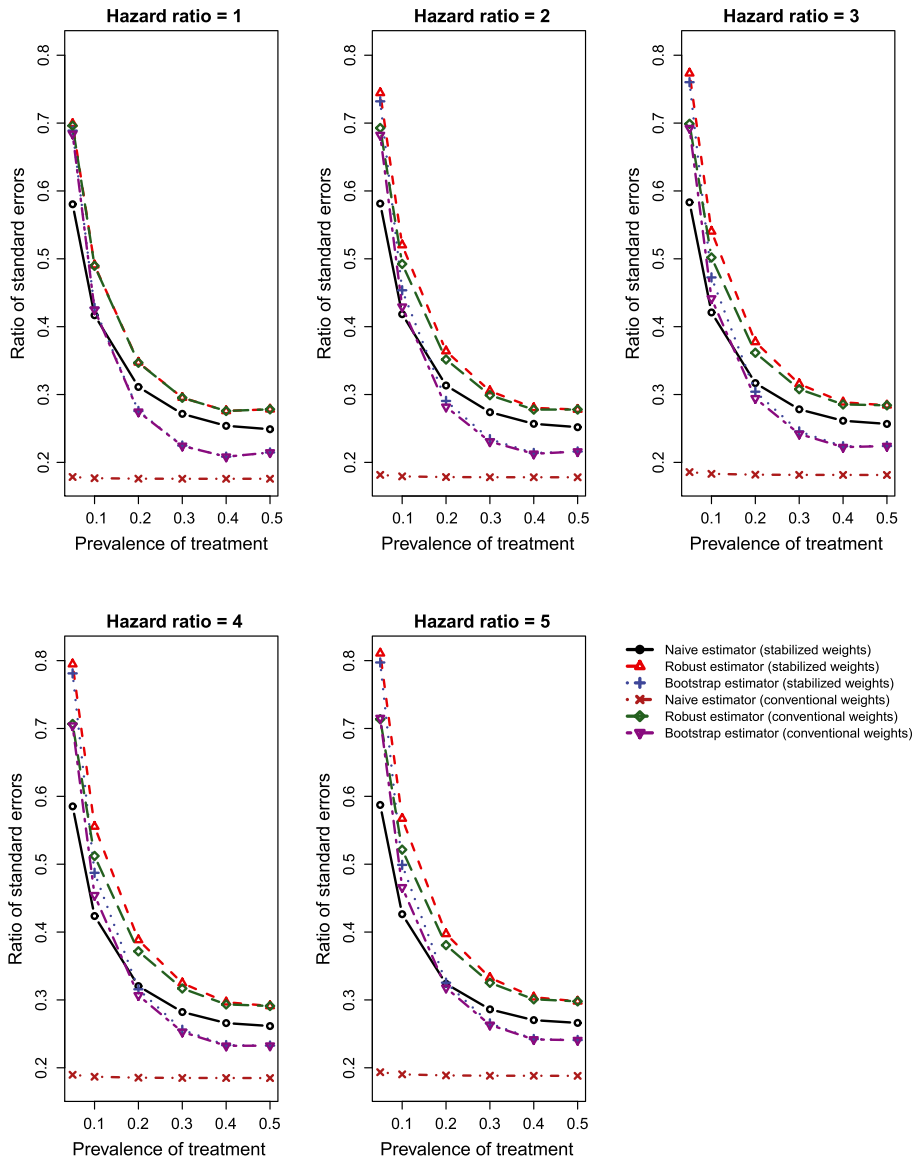

Figure 3. Mean length of estimated 95% confidence intervals: ATE.

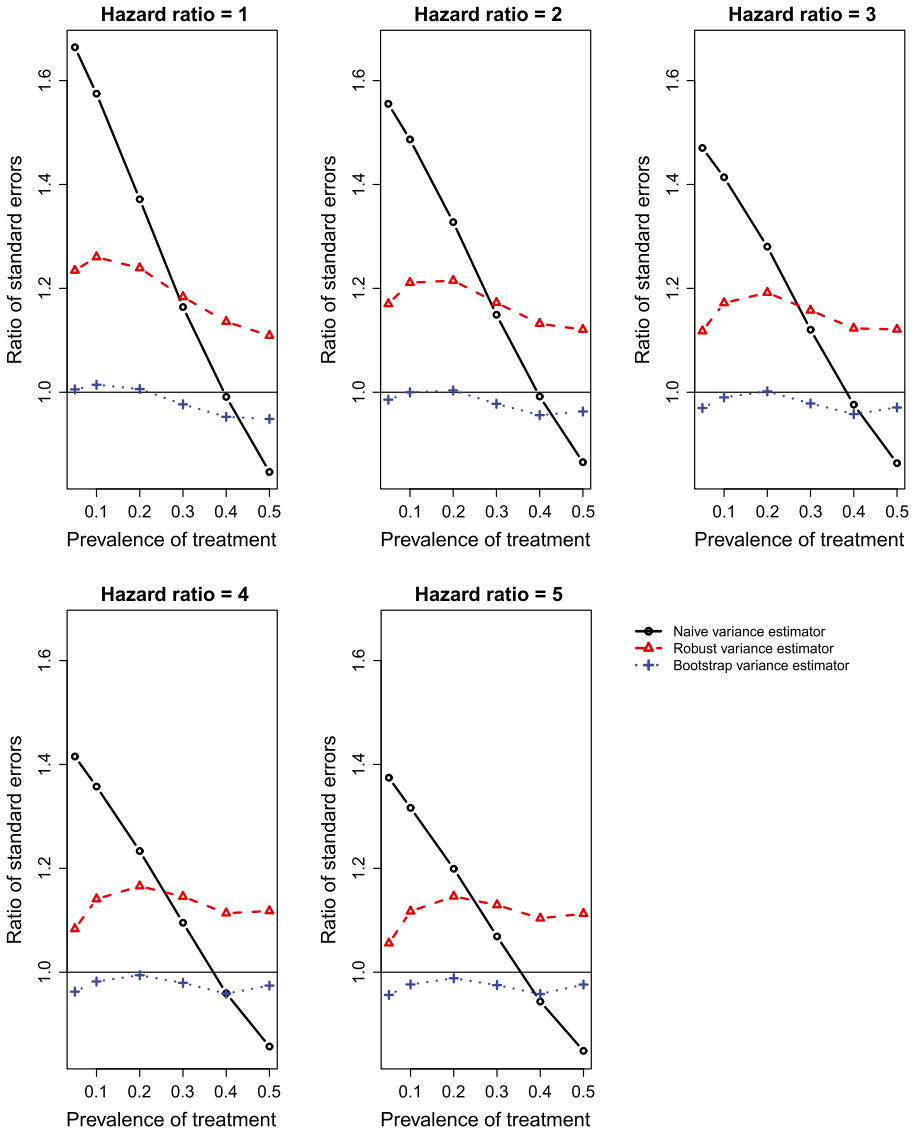

Figure 4. Ratio of mean estimated standard error to empirical standard of sampling distribution: ATT.

estimator tended to result in estimates of standard error that closely approximated the standard deviation of the empirical sampling distribution (Figure 4). In contrast to this, the naïve variance estimator and the robust variance estimator tended to result in inaccurate estimation of the standard error. Importantly, the robust variance estimator systematically over-estimated the sampling variability of the estimated log-hazard ratio. The use of the bootstrap estimator tended to result in estimated 95% confidence intervals with approximately correct coverage rates (Figure 5). The use of the other two variance estimators tended to result in 95% confidence intervals whose coverage rates differed from the nominal level. In examining the upper left panel of Figure 5, one observes that the bootstrap estimator tended to have an approximately correct type I error rate, while the other two methods tended to have incorrect type I error rates. Finally, the use of the bootstrap estimator tended to result in the narrowest 95% confidence intervals (Figure 6).

6. Discussion

We conducted an extensive series of Monte Carlo simulations to examine the performance of different variance estimators when using IPTW with the propensity score to estimate the effect of treatment on survival outcomes. We briefly summarize our findings and place them in the context of the existing literature. We found that, when estimating either the ATE or the ATT, a bootstrap-based estimator accurately estimated the sampling variability of the log-hazard ratio and tended to result in estimated

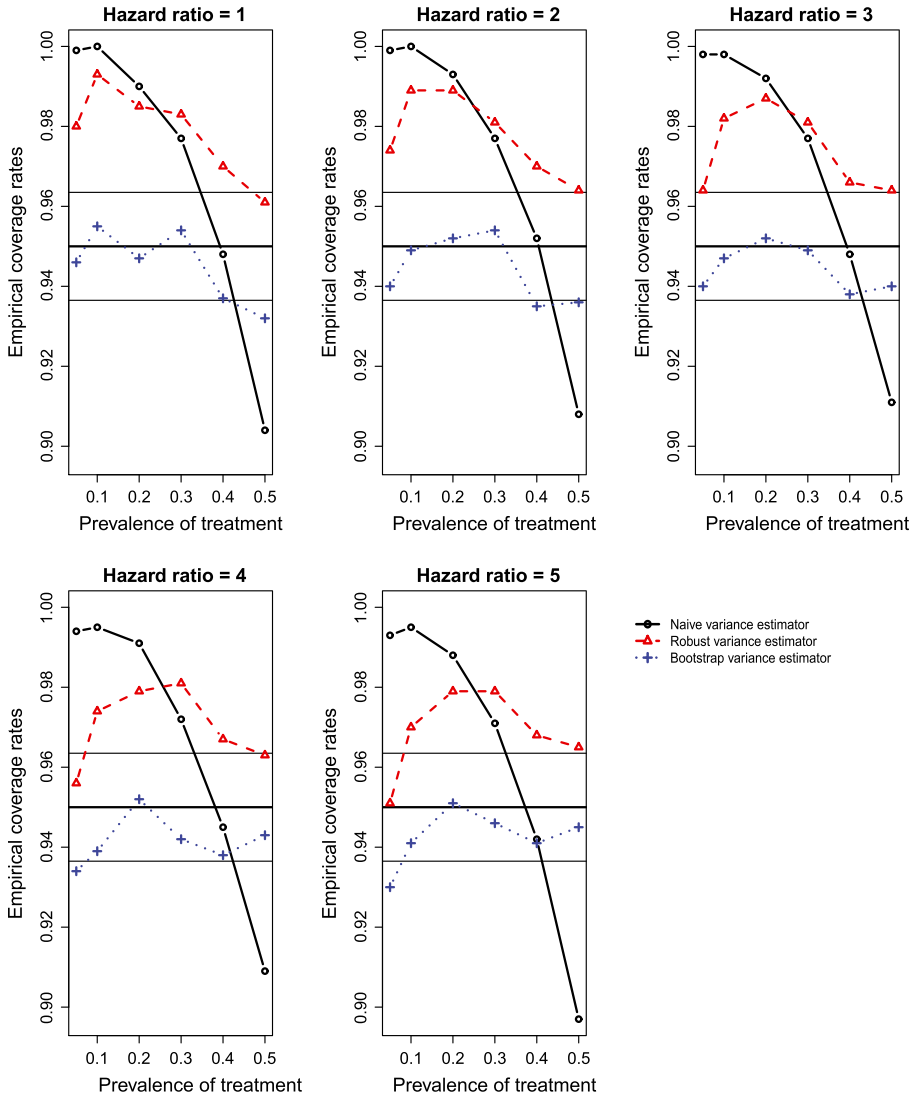

Figure 5. Empirical coverage rates of estimated 95% confidence intervals: ATT.

confidence intervals that had approximately correct coverage rates. In contrast to this, the use of a robust variance estimator tended to result in estimated standard errors that were biased upwards, regardless of which set of weights was used. Furthermore, the use of naïve variance estimator with the conventional IPTW-ATE weights resulted in estimates of standard error that were biased downwards. Finally, the bootstrap estimators tended to have an approximately correct type I error rates, while the other methods tended to have incorrect type I error rates.

Based on the observations from our Monte Carlo simulations, we would encourage analysts always to use the bootstrap estimate of the standard error rather than using the naïve or robust variance estimator provided by statistical software packages. The naïve variance estimator with the conventional IPTW weights resulted in substantial bias in estimating the standard error across all simulation settings. Similarly, the robust variance estimator (with either the conventional IPTW weights or the stabilized weights) tended to result in greater bias compared to the use of the bootstrap. In some settings, the use of a naïve variance estimator with the stabilized weights resulted in estimates of standard error that were within 10% of the true standard deviation of the sampling distribution. This tended to occur when the hazard ratio was large (hazard ratio of 4 or 5) and the prevalence of treatment was moderate (0.2 to 0.5). However, even in these scenarios, it was preferable to use the bootstrap estimator. For these reasons, we recommend that the bootstrap estimator be used in all settings.

Our findings regarding the use of the naïve variance estimator when using the IPTW-ATE weights were not surprising. As noted above, Hernan et al. note that weighting induces a within-subject correlation in outcomes in the weighted sample [23]. Thus, the assumption of independence required by the

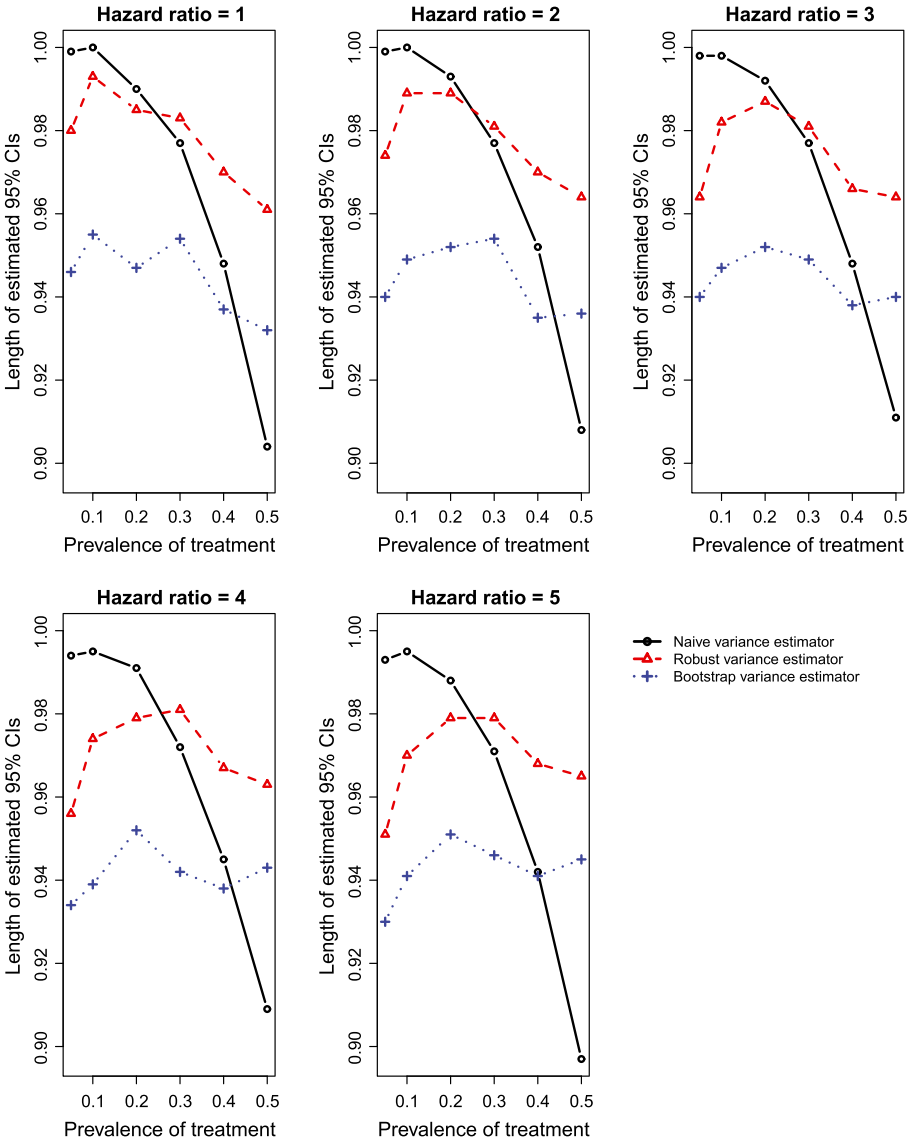

Figure 6. Mean length of estimated 95% confidence intervals: ATT.

maximum partial likelihood estimator is violated. In our simulation settings, ignoring the weighting induced an artificial inflation in the sample size which results in an artificially low estimate of the standard error of the estimated log-hazard ratio. Thus, across all scenarios, we observed that the mean estimated standard error was substantially smaller than the empirical standard deviation of the sampling distribution of the estimated log-hazard ratios.

Methods for variance estimation when using propensity score methods have received only minor attention in the methodological literature. In the context of propensity score matching, studies using Monte Carlo simulations have found that variance estimation needs to account for the matched nature of the sample and that naïve estimates of variance are inappropriate [7,8,10,11]. When using propensity-score matching *with* replacement, Abadie and Imbens found that the use of bootstrapping to estimate standard errors was inappropriate [9]. However, when matching on the propensity score *without* replacement, Austin and Small found that bootstrapping performed well [12]. When using IPTW with the propensity score to estimate linear treatment effects (i.e. differences in means or differences in proportions), Lunceford and Davidian derived formal variance estimators [13]. When fitting a regression model in a sample weighted using the IPTW weights, Joffe et al. recommended using a robust variance estimator [21]. The current study is, to the best of our knowledge, the first to comprehensively examine variance estimation when using IPTW to estimate the effect of treatment on survival outcomes. In a literature review reported in a recent article describing diagnostics for use with IPTW, it was observed that the frequency with which IPTW is being used is increasing rapidly over time [32]. Given

the increasing popularity of IPTW-based methods and the frequency with which survival outcomes occur in the medical literature [6], the results described in the current study will likely be of interest to a wide body of researchers and applied analysts.

A paper by Xu et al. is particularly relevant to the current study [25]. They used Monte Carlo simulations to compare three estimation methods (conventional IPTW-ATE weights with a naïve variance estimator vs. stabilized IPTW-ATE weights with a naïve variance estimator vs. conventional IPTW-ATE weights with a robust variance estimator) when using a logistic regression model to estimate the effect of treatment on a binary outcome. Their primary focus was on two issues: the size of the weighted sample and type I error rates. They found that the use of conventional IPTW-ATE weights tended to result in a weighted sample that was approximately double the size of the unweighted sample, while the use of stabilized IPTW-ATE weights tended to result in a weighted sample that was approximately the same size as the unweighted sample. Consequently, the use of conventional IPTW-ATE weights with a naïve variance estimator resulted in inflated type I error rates. However, the use of stabilized IPTW-ATE weights with a naïve variance estimator tended to result in approximately correct type I error rates. Finally, they found that the use of stabilized IPTW-ATE weights with a robust variance estimator resulted in type I error below lower than the advertised rate of 5%. Their finding that the use of conventional IPTW-ATE weights with a naïve variance estimator resulted in inflated type I error rates is similar to our finding that this estimator resulted in variance estimates that were too small (Figure 1) and that type I error rates were substantially inflated (top left pane of Figure 2). However, we also found that the use of stabilized IPTW-ATE weights with a naïve variance estimator tended to result in biased estimation of variance, although to a lesser degree than was observed for the conventional IPTW-ATE weights. These differences may be attributable to differences in the nature of the outcome (binary vs. time-to-event) or the simulation design (their simulation design used either a single binary covariate or a binary covariate and a continuous covariate while our design used 10 baseline covariates that were a mixture of continuous and dichotomous).

A potential explanation for the sub-optimal performance of the robust variance estimator is provided by a study by Williamson et al. that examined variance reduction in randomized trials by using IPTW using the propensity score [33]. When considering continuous or binary outcomes, they derived a variance estimator for the appropriate measure of effect (difference of means for continuous outcomes; risk difference, relative risk and odds ratio for binary outcomes) that accounted for the fact that the weights had been estimated, rather than known with certainty. As mentioned above, Hernan et al. noted that weighting induced a within-subject correlation in outcomes and that the use of robust variance estimator was encouraged to address lack of independence in the weighted sample [23]. However, while the robust variance estimator accounts for the lack of independence, it does not account for the fact that the propensity score was estimated rather than known with certainty. In contrast to this, the use of bootstrapping implicitly accounts for the estimation of the propensity score as the propensity score is estimated within each bootstrap sample. Thus, the bootstrap estimate of the standard error incorporates sampling variability in the estimated propensity score. Unfortunately, the variance estimators proposed by Williamson et al. have not been extended to the case of survival outcomes. In subsequent research it would be important to extend the variance estimator of Williamson et al. for use with a weighted Cox regression model and to compare its performance to that of the methods considered in the current study. We suspect that its performance would be similar to that of bootstrapping, as both methods account for the estimation of the propensity score. However, such a derivation is beyond the scope of the current study.

We began our analyses by conducting empirical analyses using patients discharged from hospital with a diagnosis of AMI. In this case study, we observed that the estimated standard error obtained using the naïve variance estimator with the conventional IPTW-ATE weights was substantially smaller than those obtained using either the robust variance estimator or the bootstrap estimator. This observation reflects one of the findings of the subsequent Monte Carlo simulations in which we found that the naïve variance estimator with the conventional IPTW-ATE weights resulted in substantial under-estimation of the true sampling variability. While the case study was intended to illustrate the application of the different methods, it also served a secondary purpose. The analysis of the case study data provided parameter estimates that informed the design of the subsequent Monte Carlo simulations. Thus, the simulated data reflected the data used in the case study, leading the simulations to reflect a realistic clinical scenario. As such, our simulations were similar to the plasmode simulations described by Franklin et al. [34].

In summary, the current study was motivated by prior research that observed that the use of a robust sandwich-type variance estimator when using IPTW to estimate the effect of treatment on survival outcomes resulted in biased estimates of standard errors and confidence intervals with incorrect coverage

rates [8]. Based on the results of the current study, we recommend that the bootstrap be used to estimate standard errors and confidence intervals when fitting a Cox proportional hazards model in a sample weighted using the IPTW.

Appendix

Statistical software code for fitting a weighted Cox regression model with robust standard errors

```
# R code for fitting a weighted Cox model with robust standard errors
coxph(Surv(time,status) ~ treat + cluster(id),weights = wt.ate,data = dataset)

# time is the variable denoting survival time.
# status is the indicator variable denoting event status (1 = event occurred; 0 = subject was
# censored).
# treat is the binary indicator variable denoting treatment status (1 = treated; 0 = control).
# id is the identification variable that identifies unique subjects.
# wt.ate is the variable containing the inverse probability of treatment weights.
# dataset is the data frame containing the data

/* SAS code for fitting a weighted Cox model with robust standard errors */
Proc phreg data = dataset covs(agg);
   Model time*status(0) = treat /ties = efron;
 Id id;
Run;
```

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 an operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin was supported by Career Investigator awards from the Heart and Stroke Foundation. The Enhanced Feedback for Effective Cardiac Treatment (EFFECT) data used in the study was funded by a CIHR Team Grant in Cardiovascular Outcomes Research. These datasets were linked using unique, encoded identifiers and analyzed at the Institute for Clinical Evaluative Sciences (ICES). This study was approved by the institutional review board at Sunnybrook Health Sciences Centre, Toronto, Canada.

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 tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivariate Behavioral Research* 2011; **46**:119–151.
3. Austin PC. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 2011; **46**:399–424.
4. 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.
5. 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.
6. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology* 2010; **63**(2):142–153.
7. 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. The performance of different propensity score methods for estimating marginal hazard ratios. *Stastisics in Medicine* 2013; **32**(16):2837–2849.
9. Abadie A, Imbens GW. Notes and comments on the failure of the bootstrap for matching estimators. *Econometrica* 2008; **76**(6):1537–1557.
10. Austin PC, Type I, Rates E. Coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *International Journal of Biostatistics* 2009; **5**(1): Article 13. DOI: 10.2202/1557-4679.1146.

11. 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.
12. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statisics in Medicine* 2014; **33**(24):4306–4319.
13. 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.
14. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Statistics in Medicine* 2010; **29**(3):337–346.
15. 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.
16. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 2004; **9**(4):403–425.
17. 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.
18. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine* 2004; **75**:45–49.
19. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**(6):656–664.
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. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *Journal of the American Statistical Association* 1989; **84**(408):1074–1078.
23. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 2000; **11**(5):561–570.
24. van der Wal WM, Geskus RB. ipw: an R package for inverse probability weighting. *Journal of Statistical Software* 2011; **43**(13). https://www.jstatsoft.org/article/view/v043i13.
25. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value in Health* 2010; **13**(2):273–277.
26. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stastisics in Medicine* 2014; **33**(7):1242–1258.
27. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine* 2006; **25**(12):2084–2106.
28. Tu JV, Donovan LR, Lee DS, Wang JT, Austin PC, Alter DA, Ko DT. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Journal of the American Medical Association* 2009; **302**(21):2330–2337.
29. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall: New York, NY, 1993.
30. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723.
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, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine* 2015; **34**(28):3661–3679.
33. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* 2014; **33**(5):721–737.
34. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis* 2014; **72**:219–226.

