Received: 6 March 2019 | Revised: 13 January 2020 | Accepted: 21 January 2020

DOI: 10.1002/sim.8502

RESEARCH ARTICLE

Statistics in Medicine WILEY

Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes

Peter C. Austin[1,2,3] | **Guy Cafri**[4]

[1] ICES, Toronto, Ontario, Canada

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

[3] Sunnybrook Research Institute, Toronto, Ontario, Canada

[4] Johnson & Johnson Medical Devices, San Diego, California

Correspondence
Peter C. Austin, ICES G106, 2075 Bayview Avenue, Toronto, Ontario M4N 3M5, Canada.
Email: peter.austin@ices.on.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Number: MOP 86508; Heart and Stroke Foundation of Canada, Grant/Award Number: Mid-Career Investigator Award

Propensity-score matching is a popular analytic method to estimate the effects of treatments when using observational data. Matching on the propensity score typically requires a pool of potential controls that is larger than the number of treated or exposed subjects. The most common approach to matching on the propensity score is matching without replacement, in which each control subject is matched to at most one treated subject. Failure to find a matched control for each treated subject can lead to "bias due to incomplete matching." To avoid this bias, it is important to identify a matched control subject for each treated subject. An alternative to matching without replacement is matching with replacement, in which control subjects are allowed to be matched to multiple treated subjects. A limitation to the use of matching with replacement is that variance estimation must account for both the matched nature of the sample and for some control subjects being included in multiple matched sets. While a variance estimator has been proposed for when outcomes are continuous, no such estimator has been proposed for use with time-to-event outcomes, which are common in medical and epidemiological research. We propose a variance estimator for the hazard ratio when matching with replacement. We conducted a series of Monte Carlo simulations to examine the performance of this estimator. We illustrate the utility of matching with replacement to estimate the effect of smoking cessation counseling on survival in smokers discharged from hospital with a heart attack.

KEYWORDS

matching, Monte Carlo simulations, observational study, propensity score, survival analysis.

1 | INTRODUCTION

Observational studies are increasingly being used to estimate the effects of treatments, interventions, and exposures on outcomes. Due to nonrandom treatment assignment, treated subjects often differ systematically at baseline from control subjects. These systematic baseline differences between treatment groups in observational studies result in treatment assignment being confounded with baseline characteristics. Consequently, statistical methods must be used to remove or minimize the effect of this confounding so that valid inferences on the effect of treatment can be drawn from observational studies.

Propensity score methods are increasingly being used to minimize confounding in observational studies examining the effect of treatment on outcomes. The propensity score is the probability of treatment assignment conditional on measured baseline covariates.[1] 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 using the propensity score, and covariate adjustment using the propensity score.[1,2] Propensity score matching is frequently used in the medical literature.[3-5]

There are a large number of different matching algorithms that have been used when matching on the propensity score: pair matching, many-to-one matching, variable ratio matching, full matching, greedy nearest neighbor matching (NNM), caliper matching, optimal matching, matching with replacement and matching without replacement (note that a given implementation of matching can combine multiple elements from this list).[6-11] The most common implementation of propensity-score matching appears to be greedy NNM without replacement (with or without calipers). A limitation of most matching algorithms (with the exception of full matching) is that they require a pool of potential controls that is substantially larger than the number of treated subjects. The reason for this is that matching on the propensity score estimates the average treatment effect in the treated (ATT).[12,13] Failure to identify a matched control subject can result in "bias due to incomplete matching."[6] This bias arises because one is attempting to estimate the effect of treatment in treated subjects by using a subset of the treated subjects. Frequently this subset of matched treated subjects differs systematically from the population of treated subjects (eg, treated subjects with higher propensity scores are more likely to be excluded from the matched sample because there are insufficient control subjects with high propensity scores). The potential for this bias limits the utility of matching without replacement on the propensity score in settings in which the pool of potential controls is smaller than the sample of treated subjects. Matching with replacement is an attractive approach that would allow analysts to potentially circumvent biases arising from incomplete matching. By allowing control subjects to be matched to multiple treated subjects, the analyst increases the likelihood that a high-quality match can be identified for each treated subject.

An important issue when using propensity-score matching is how to estimate the variance (or standard error (SE)) of the estimated treated effect. Some authors have suggested that the propensity score matched sample can be treated as consisting of independent subjects and that conventional statistical methods can be used to compare outcomes between treatment groups.[14] However, more recent studies have shown that it is preferable to account for the matched nature of the study when estimating the variance of the estimated treatment effect.[15-18] Ideally, a variance estimator would account for the following issues: (a) the variability induced by the matching process; (b) the true propensity score is estimated and not known with certainty; (c) the propensity score may have been estimated using data-adaptive methods; and (d) a within-matched pair correlation in outcomes may be induced by matching. Abadie and Imbens[19] derived a formal variance estimator for matched estimators when outcomes are continuous. This estimator accounted for the variability introduced by the matching procedure, but not for other sources of variability. In the absence of formal variance estimators for other measures of effect, investigators have examined whether variance estimators derived in other contexts can be applied to samples constructed using propensity score matching. For example, it has been shown that McNemar's test can be used to compare differences in proportions in propensity-score matched samples.[16] Similarly, the robust variance estimator proposed by Lin and Wei has been found to perform well when matching without replacement on the propensity and estimating marginal hazard ratios.[17,20]

While it is a potentially attractive strategy, matching with replacement is rarely used in the medical and epidemiological literature. One reason for the infrequency with which matching with replacement is used is that methods for variance estimation for the resultant treatment effect have not been described for measures of effect that are commonly used in the medical and epidemiological literature. When matching on the propensity score, a within-matched set correlation in outcomes has been induced, which needs to be accounted for when estimating the variance of the treatment effect.[15-18] When matching with replacement one also needs to account for the fact that the same control subject has potentially been matched to multiple treated subjects. Hill and Reiter[21] examined variance estimation when estimating linear treatment effects for continuous outcomes. In medical and epidemiological research, binary and time-to-event outcomes are common. However, variance estimation for common measures of effect such as the relative risk, risk difference, and hazard ratio has not been described when using propensity score matching with replacement. The use of an appropriate variance estimator is essential to statistical inference. Abadie and Imbens[22] have demonstrated that bootstrap methods may not be used when matching on the propensity score *with* replacement. Given the lack of a variance estimator for common measures of effect such as the hazard ratio and the inability to use the bootstrap when matching with replacement, there is a need for a variance estimator to be proposed and evaluated. We propose to modify a variance estimator for use with clustered data in which there are two sources of clustering and to examine the performance of this modified estimator when estimating hazard ratios using matching with replacement.

The objective of the current study was to examine the performance of propensity score matching with replacement to estimate marginal hazard ratios when outcomes are time-to-event in nature. The article is structured as follows. In Section 2, we review the use of propensity score matching with survival outcomes and propose a variance estimator for the marginal log-hazard ratio when using propensity score matching with replacement. In Section 3, we describe the design of an extensive set of Monte Carlo simulations to examine the performance of this variance estimator. We compare the performance of the proposed estimator to two alternative estimators. In Section 4, we report the results of these simulations. In Section 5, we provide a case study in which we illustrate the utility of matching with replacement. In Section 6, we summarize our findings and place them in the context of the existing literature.

2 | PROPENSITY SCORE MATCHING AND SURVIVAL OUTCOMES

2.1 | Previous research on propensity score matching and survival outcomes

Previous studies have demonstrated that pair-matching on the propensity score when matching without replacement leads to biased estimation of conditional hazard ratios, but unbiased estimation of marginal hazard ratios.[17,23] Estimation of the marginal hazard ratio is achieved by using a univariate Cox proportional hazards regression model in the matched sample to regress the hazard of the outcome on an indicator variable denoting treatment status. A robust, sandwich estimate that accounts for the clustering within matched sets was shown to result in appropriate estimates of the SE of the log-hazard ratio.[17] While less frequently used, marginal hazard ratios can be estimated using full matching on the propensity score.[24,25]

2.2 | Variance estimation when matching with replacement

Matching with replacement induces two types of correlations that must be accounted for when estimating the variance of estimated treatment effects. The first is a within-matched set correlation in outcomes. Matched subjects within the same matched set have similar values of the propensity score. Subjects who have the same value of the propensity score have measured baseline covariates that come from the same multivariate distribution.[1] In the presence of confounding, baseline covariates are related to the outcome. Thus, matched subjects are more likely to have similar outcomes compared to two randomly selected subjects. The second source of correlation is induced by repeated use of control subjects. Failure to account for this correlation and acting as though the matched control subjects were independent observations will likely result in estimated standard errors that are artificially small and estimated confidence intervals that are artificially narrow. Added complexity is introduced by having subjects cross-classified with matched sets such that the same control subject can belong to more than one matched set.

When matching without replacement, as noted above, it has been shown that fitting a Cox proportional hazards model and using a robust sandwich-type variance estimator allows for accurate estimation of the sampling variance of the estimated log-hazard ratio.[17] Our proposed variance estimator for use with matching with replacement is motivated by a variance estimator proposed by Miglioretti and Heagerty for use with generalized linear models estimated using generalized estimation equation (GEE) methods when there are multiple nonnested sources of clustering.[26] Our proposed method involves fitting a Cox proportional model to the matched sample and obtaining three different variance estimators for the log-hazard ratio. The conventional univariate proportional hazards model is fit to the matched sample: $\log(h) = \beta Z$, where Z is an indicator variable denoting treatment status ($Z = 1$ for active treatment of interest vs $Z = 0$ for the control treatment) and β denotes the log-hazard ratio. The first variance estimator (V_1) for the log-hazard ratio is the robust variance estimator proposed by Lin and Wei that accounts for within matched-pair correlation in outcomes[20] (Miglioretti and Heagerty, working with GEEs in the context of generalized linear models, proposed the use of an independence working correlation structure). Thus, subjects in the same matched set are considered as having outcomes that are correlated with one another, while subjects from different matched sets are considered independent of one another. The second variance estimator (V_2) for the log-hazard ratio is the robust sandwich-type variance estimator proposed by Lin and Wei, which accounts for clustering within subjects. Thus, outcomes are considered independent between distinct subjects. The third variance estimator (V_3) for the log-hazard ratio is a robust sandwich-type variance estimator that accounts for clustering in the cross-classification of the two sources of clustering. The data are cross-classified because controls can belong to more than one matched set. However, because each subject can appear in at most one matched set, this reduces (in our application) to having maximum cluster sizes equal to one upon cross-classification, which results in independence. Let K_1 denote the number of matched pairs, K_2 denote the number of number of unique individuals in the matched sample, and K_3 denote the overall size of the matched sample ($K_3 = 2K_1$). Then the proposed variance estimator for the estimated log-hazard ratios is $\frac{K_1}{K_1-1}V_1 + \frac{K_2}{K_2-1}V_2 - \frac{K_3}{K_3-1}V_3$. The correction factor for each variance term is used to correct for bias in the variance estimate resulting from potentially having a small number of clusters.[27] Note that if matching without replacement (or if no control was selected for inclusion in more than one matched set), this would reduce to V_1, which is the estimator that has already been proposed for matching *without* replacement. R code for implementing this estimator is described in Data S1.

3 | MONTE CARLO SIMULATIONS

We used a series of Monte Carlo simulations to examine the performance of the proposed variance estimator for the estimate of the marginal hazard ratio obtained when using propensity-score matching with replacement. The design of these simulations was similar to that of a series of simulations used to compare the performance of different propensity score methods for estimating marginal hazard ratios.[17] We will compare the proposed variance estimator for the estimated marginal hazard ratio with two alternative variance estimators.

3.1 | Data-generating process

We simulated data for a large superpopulation consisting of 1 000 000 subjects. For each subject, we simulated ten baseline covariates (X_1–X_{10}) from independent standard normal distributions. Of these ten covariates, seven affected treatment selection (X_1, X_2, X_4, X_5, X_7, X_9, X_{10}), while seven affected the outcome (X_2, X_3, X_5, X_6, X_8, X_9, 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: $\text{logit}(p_i) = \alpha_{0,\text{ treat}} + \alpha_W x_1 + \alpha_W x_2 + \alpha_M x_4 + \alpha_M x_5 + \alpha_S x_7 + \alpha_S x_8 + \alpha_{VS} x_{10}$. An iterative bisection approach was used to determine the value of the intercept of the treatment-selection model ($\alpha_{0,\text{treat}}$) so that the proportion of subjects in the superpopulation that were treated was fixed at the desired proportion (see later for factors allowed to vary in the simulations). The regression coefficients α_W, α_M, α_S, α_{VS} were set to log(1.25), log(1.5), log(1.75), and log (2), respectively. These were intended to denote weak, moderate, strong, and very strong treatment-assignment affects. For each subject, treatment status was generated from a Bernoulli distribution with subject-specific parameter p_i.

For each subject in the large superpopulation, we generated two potential time-to-event outcomes: the potential outcome under control and the potential outcome under treatment. For each potential outcome, we used a data-generating process for time-to-event outcomes described by Bender et al.[28] When simulating the potential outcome under control, the linear predictor was defined as $\text{LP} = \alpha_W x_2 + \alpha_W x_3 + \alpha_M x_5 + \alpha_M x_6 + \alpha_S x_8 + \alpha_S x_9 + \alpha_{VS} x_{10}$. For each subject, we generated a random number from a standard uniform distribution: $u \sim U(0,1)$. A survival or event time was generated for each subjects as follows: $\frac{-\log(u)}{(\lambda e^{\text{LP}})^{1/\eta}}$. We set λ and η to be equal to 0.00002 and 2, respectively, as has been done in previous studies.[17,24,25,29-31] When simulating the potential outcome under treatment, the linear predictor was defined as $\text{LP} = \beta_{\text{treat}} + \alpha_W x_2 + \alpha_W x_3 + \alpha_M x_5 + \alpha_M x_6 + \alpha_S x_8 + \alpha_S x_9 + \alpha_{VS} x_{10}$, while the other computations were unchanged.

This process for generating outcomes employs a conditional model and results in data with a specified conditional treatment effect. We wanted to generate data in which there was a specified marginal hazard ratio. The underlying marginal hazard ratio when the ATT is the target estimand can be determined as follows: The sample is restricted to those subjects who were treated (since the target estimand is the ATT). Then using both potential outcomes, we regressed the potential outcomes on treatment status (thus, each treated subject contributed two outcomes to the analysis, one under each treatment condition) and the marginal hazard ratio was estimated. We used an iterative bisection process to determine the value of β_{treat} (the conditional log-hazard ratio) that induced the desired marginal hazard ratio. This process is based on methods that have been used in previous studies to induce data with a desired marginal hazard ratio.[17,24,25,29-32] We thus simulated data for a large super-population such that the prevalence of treatment is set at a fixed value and the underlying marginal hazard ratio for treatment is set at a fixed value.

3.2 | Analyses in simulated data

From the large super-population of size 1 000 000, we drew a random sample of size 1000. In this sample of size 1000, we estimated the propensity score by using a logistic regression model to regress treatment status on the seven variables that are associated with the outcome, as this has been shown to be a good strategy for variable selection for the propensity score model.[33] Using the logit of the estimated propensity score, treated subjects were matched to control subjects using two different matching algorithms: (a) NNM with replacement and (b) NNM with replacement with a caliper restriction equal to 0.2 of the SD of the logit of the propensity score.[34] Within each of the two matched samples, we determined the proportion of controls that were included more than once. We refer to this as the proportion of controls that were recycled (or used multiple times). This proportion can range from 0 (if no control was used more than once) to $(N-1)/N$ (if the same control was matched to all of the treated subjects [N denotes the number of matched pairs]). Within each of the two matched samples, the hazard of the outcome was regressed on a single indicator variable denoting treatment status. The SE of the estimated log-hazard ratio was obtained using the estimator described in Section 2. This process was repeated 1000 times.

For comparative purposes, we considered two additional variance estimators. First, we used an independent variance estimator that assumed that all included subjects were independent of another. This is the conventional model-based estimator from the Cox proportional hazards model. Second, we used a variance estimator that accounted for clustering within matched pairs, but that ignored the fact that the same control could be included in multiple matched sets. This is the estimator that we have described as V_1 above, and which was explored elsewhere in the context of matching without replacement.[17]

We determined the mean proportion of recycled controls across the 1000 simulated datasets and for the two matching methods. In each of the 1000 simulated datasets and for each of the two matching methods, we obtained an estimate of the log-hazard ratio and its estimated SE. Let θ_i denote the estimated log-hazard ratio obtained from the ith simulated dataset, while θ denotes the true log-marginal hazard ratio. We estimated the mean treatment effect (on the log-hazard scale) as $\overline{\theta} = \frac{1}{1000}\sum_{i=1}^{1000}\theta_i$, relative bias on the hazard ratio scale was defined as $100 \times \frac{\exp(\overline{\theta})-\exp(\theta)}{\exp(\theta)}$ (we determine relative bias on the hazard ratio scale as the relative bias on the log-hazard ratio scale is not defined for a null hazard ratio). We examined the accuracy with which the estimated SEs of the estimated log-hazard ratios estimated the sampling variability of the log-hazard ratios. To do so, we compared two estimates. First, we determined the mean estimated SE of the estimated log-hazard ratio across the 1000 simulated datasets. Second, we determined the standard deviation (SD) of the estimated log-hazard ratios across the 1000 simulated datasets. We then determined the ratio of these two quantities. If the ratio equals one, then the estimated SE of the log-hazard ratio is correctly estimating the sampling variability of the estimated log-hazard ratio. Within each simulated dataset and for each of the two matching methods, we computed the 95% confidence interval for the estimated hazard ratio using normal-theory methods and the estimated SE. We then determined the proportion of 95% confidence intervals that covered the true marginal hazard ratio that was used in the data-generating process. Finally, when the true marginal hazard ratio was equal to one, we determined the proportion of samples in which the null hypothesis of a null treatment effect was rejected.

3.3 | Factors in the Monte Carlo simulations

We allowed the following factors to vary in our Monte Carlo simulations: the percentage of subjects that were treated (from 10% to 90% in increments of 10%) and the true marginal hazard ratio (1, 1.2, 1.4, 1.6, 1.8, and 2). We thus examined 54 scenarios (nine treatment prevalence × six marginal hazard ratios). We thus generated 54 superpopulations with the desired characteristics and from each superpopulation, we drew 1000 random samples in which we conducted the described statistical analyses.

3.4 | Sensitivity analysis I: Including all variables in the propensity score model

Previous research has suggested that one should include in the propensity score model either those variables that predict the outcome or that are confounders of the treatment-outcome relationship.[33,35] However, in some settings, the analyst may not have the luxury of knowing which variables are associated only with the outcome or are confounders of the treatment-outcome relationship. This set of simulations was identical to those described above, with one minor modification. When the propensity score was estimated in a simulated sample, rather than including those seven variables that were associated with the outcome, we included all 10 baseline variables.

3.5 | Sensitivity analysis II: Misspecified propensity score model

The first two sets of simulations assumed that the propensity score model had been correctly specified (either by including those variables that were associated with the outcome or by including those variables that were associated with treatment selection). In this set of simulations, we examined the effect of misspecifying the propensity score model. This set of simulations was identical to those described in Section 3.4, with one modification. When simulating treatment status, the following model was used: $\text{logit}(p_i) = \alpha_{0,\text{treat}} + \alpha_W x_1 + \alpha_W x_2 + \alpha_M x_4 + \alpha_M x_5 + \alpha_S x_7 + \alpha_S x_8 + \alpha_{VS} x_{10} + \alpha_W x_2^2 + \alpha_W x_4 x_5 + \alpha_W x_7 x_8$. Thus, the treatment-selection model included main effects for seven baseline covariates, one quadratic term, and two interactions between baseline covariates. However, when the propensity score model was estimated in each of the random samples, only main effects for the seven baseline covariates were included.

4 | MONTE CARLO SIMULATIONS—RESULTS

4.1 | Primary simulations

The proportion of matched controls that were recycled for each of the two matching methods across the different scenarios is reported in Figure 1. For each matching method, the proportion of matched controls that were recycled was proportional to the prevalence of treatment. The proportion of recycled controls was slightly higher when using NNM than when using caliper matching (differences ranged from 0.01 to 0.07 across the 54 scenarios). When using caliper matching, the proportion of recycled controls was very close to the prevalence of treatment.

The relative bias for each of the two matching methods across the different scenarios is reported in Figure 2. Relative bias was minimal across all 54 scenarios for both matching methods. Relative bias was largest when the prevalence of treatment was low and caliper matching was employed.

The accuracy with which the estimated SEs approximated the SD of the sampling distribution of the estimated log-hazard ratios is reported in Figure 3 when the true marginal hazard ratio was 1 or 2 (results for the four other

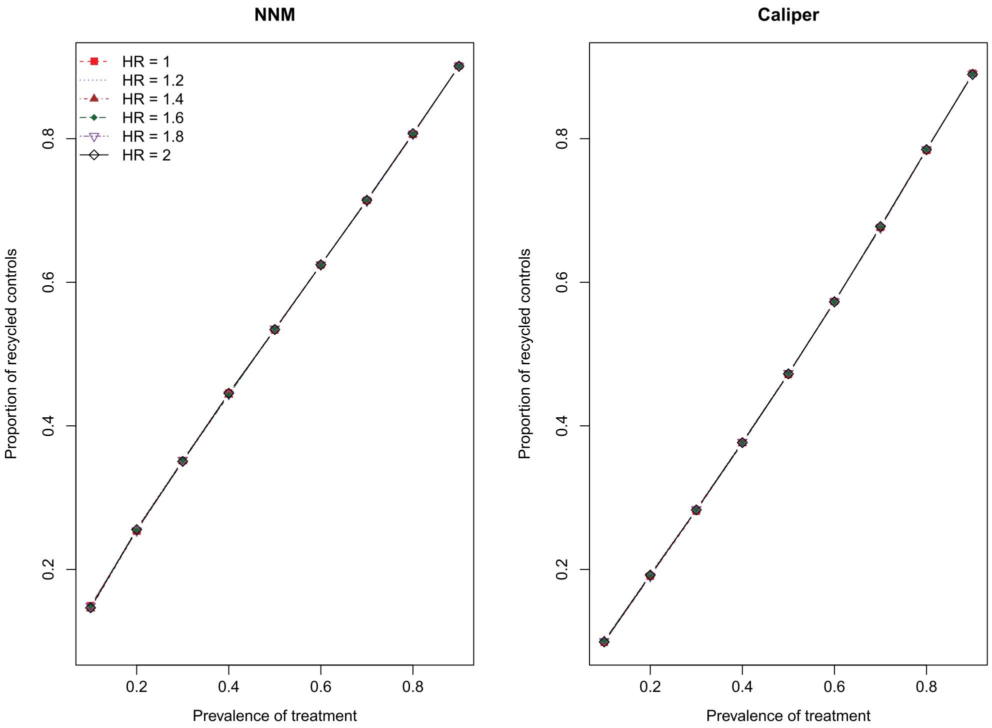


FIGURE 1 Proportion of controls rejected [Colour figure can be viewed at wileyonlinelibrary.com]

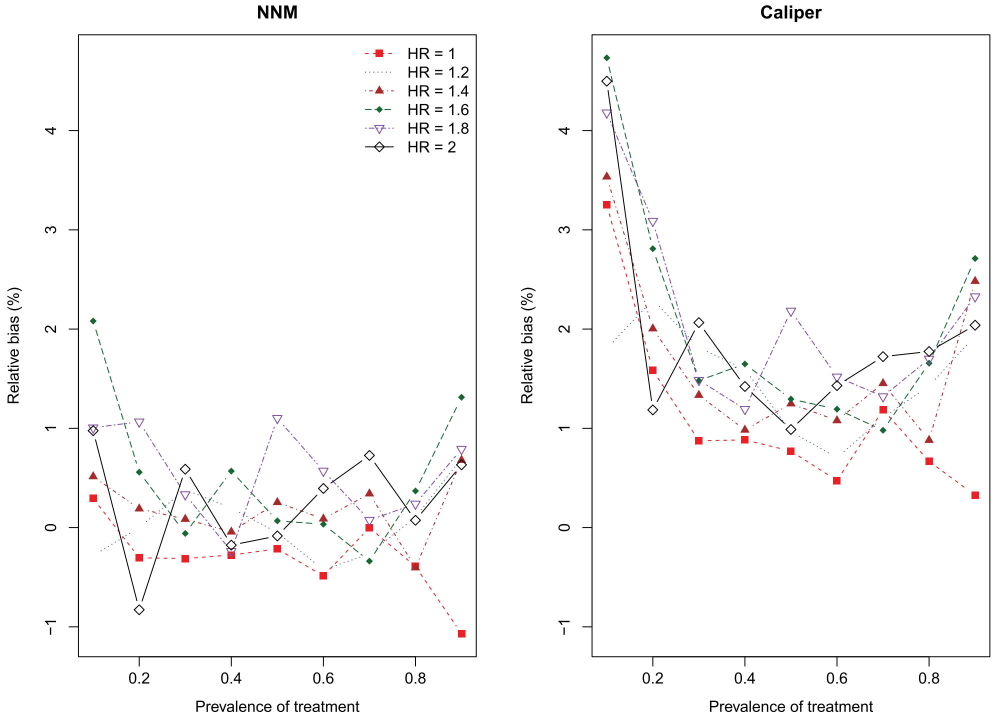


FIGURE 2 Relative bias in estimated hazard ratio [Colour figure can be viewed at wileyonlinelibrary.com]

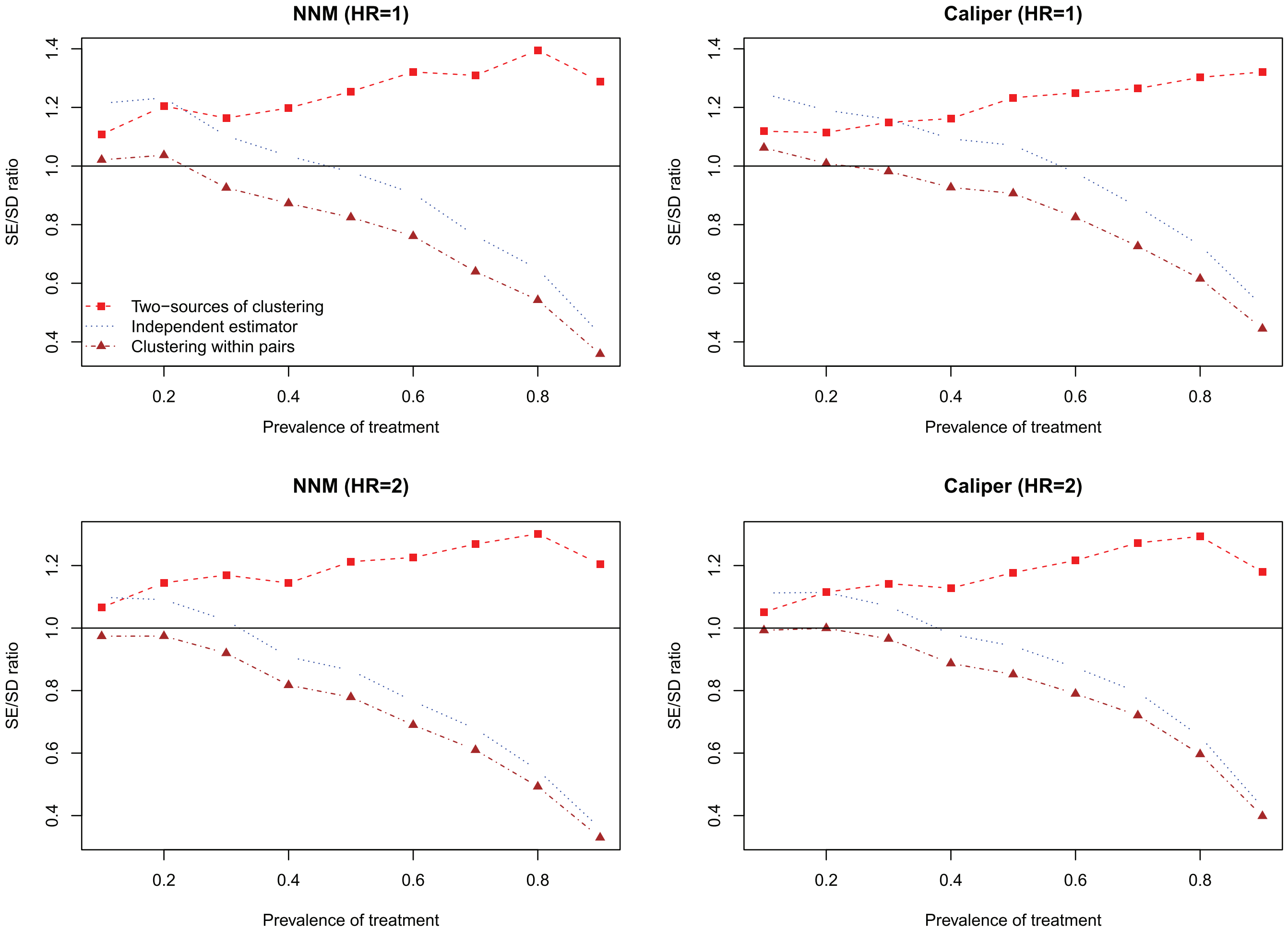


FIGURE 3 Ratio of mean estimated standard error to standard deviation of estimated log-hazard (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

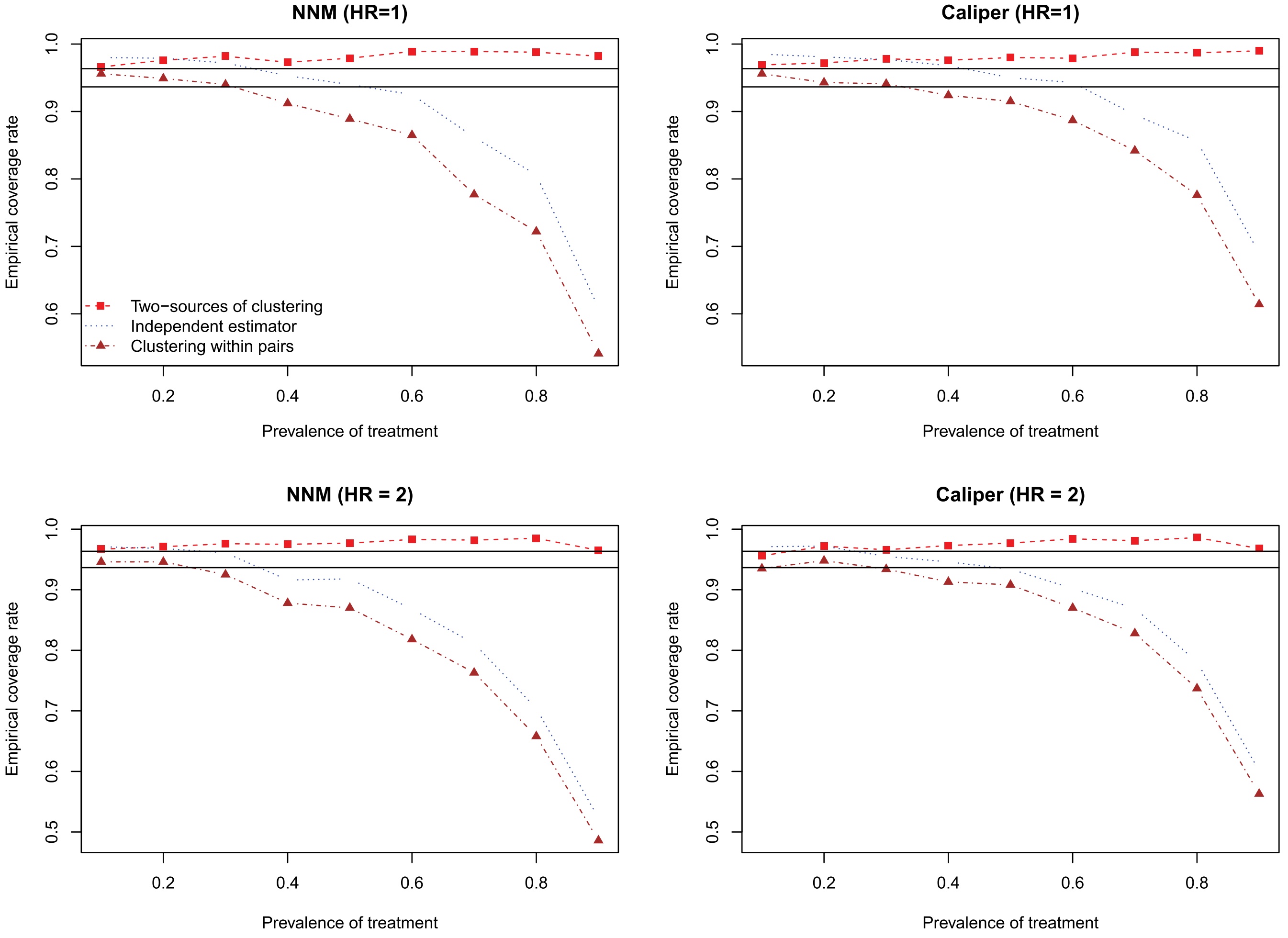


FIGURE 4 Coverage rates of estimated confidence intervals (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

hazard ratios are reported in Figures A1 to A2 in Data S2). For each of the three variance estimators, we report the ratio of the mean SE of the estimated log-hazard ratio across the 1000 simulated datasets to the SD of the estimated log-hazard ratios across the 1000 simulated datasets. Across all 54 scenarios and both matching methods, the ratio was always greater than one when using the proposed variance estimator, implying that the estimated SE of the log-hazard ratio systematically over-estimates the sampling variation of the log-hazard ratio. In general, the degree of overestimation increased as the prevalence of treatment increased (except when the prevalence of treatment was 90%, at which the ratio tended to decrease). Across the majority of scenarios, the ratio ranged from 1.07 to 1.32, denoting minor to modest overestimation of the sampling variation of the regression coefficient. When using either of the two alternative variance estimators, the estimated SE tended to underestimate the sampling variability of the log-hazard ratio, with the degree of under-estimation increasing as the prevalence of treatment increased. When the prevalence of treatment was high, the degree of underestimation was substantial.

Coverage rates of 95% confidence intervals are reported in Figure 4 for a true marginal hazard ratio of 1 or 2 (results for the other settings are reported in Figures A3 to A4 in Data S2). Due to our use of 1000 iterations per scenario in our Monte Carlo simulations, any confidence intervals whose empirical coverage rate is less than 0.9365 or greater than 0.9653 would be statistically significantly different from 0.95, using a standard normal-theory test. On each panel, we have superimposed a vertical line denoting empirical coverage rates of 0.9365 and 0.9635. When using NNM, the empirical coverage rates exceeded 0.95 across all 54 scenarios when using our proposed variance estimator. When using caliper matching, the empirical coverage rates exceeded 0.948 in all 54 scenarios when using our proposed variance estimator. When using either of the two alternative variance estimators, the empirical coverage rates of the estimated confidence intervals were substantially lower than advertised when the prevalence of treatment was high.

The empirical type I error rates when the true marginal hazard ratio was equal to one are reported in Figure 5. Across all nine scenarios and for both matching methods, the empirical type I error rates obtained using the proposed variance

FIGURE 5 Empirical type I error rate [Colour figure can be viewed at wileyonlinelibrary.com]

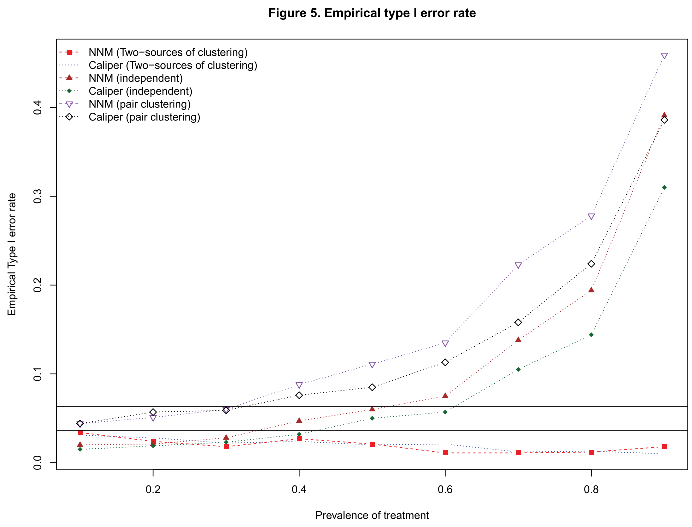
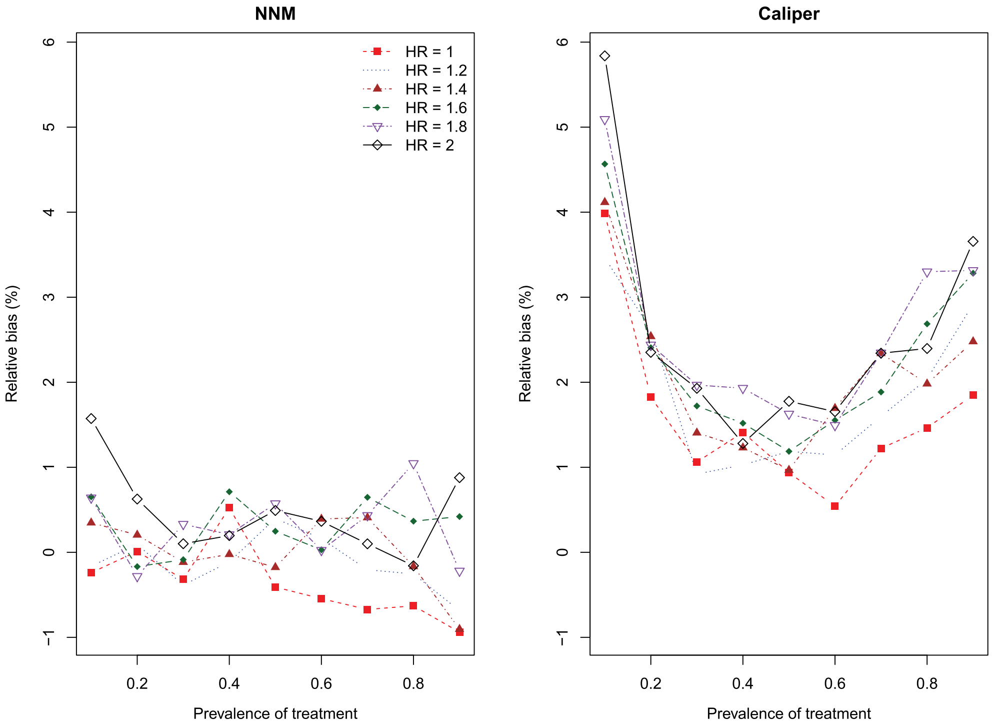


FIGURE 6 Relative bias in estimated hazard ratio [Colour figure can be viewed at wileyonlinelibrary.com]



estimator were below 0.0365, indicating that they were significantly different from the advertised level of 0.05, based on a standard normal-theory test. Thus, the variance estimator results in a statistical test that is conservative, rejecting the null hypothesis less frequently than advertised. In contrast to this, the empirical type I error rates were substantially higher than 0.05 when using alternative variance estimators, particularly when the prevalence of treatment was high.

4.2 | Sensitivity analysis I: Including all variables in the propensity score model

Results for the first sensitivity analysis are reported in Figure 6 (relative bias in estimated hazard ratio), Figure 7 (ratio of mean estimated SE to SD of sampling distribution), Figure 8 (empirical coverage rates of confidence intervals), Figure 9 (empirical type I error rates), and in Figures B1 to B4 in Data S2. The figures are structurally similarly to those for the primary set of simulations. The observed results were qualitatively similar to those obtained in the primary set of simulations.

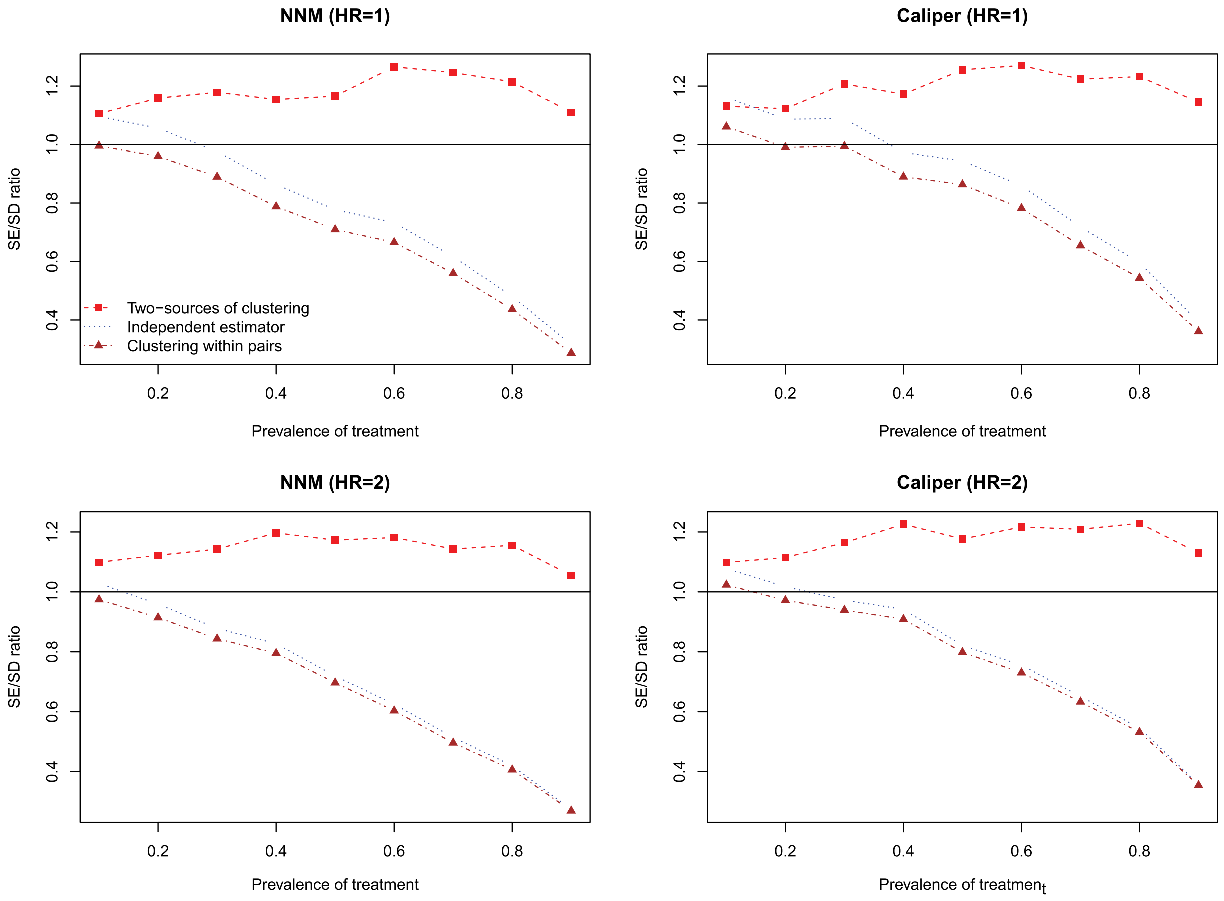


FIGURE 7 Ratio of mean estimated standard error to standard deviation of estimated log-hazard ratio (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | Sensitivity analysis II: Misspecified propensity score model

Results for the second sensitivity analysis are reported in Figure 10 (relative bias in estimated hazard ratio), Figure 11 (ratio of mean estimated SE to SD of sampling distribution), Figure 12 (empirical coverage rates of confidence intervals), Figure 13 (empirical type I error rates), and in Figures C1 to C4 in Data S2. The figures are structurally similarly to those for the primary set of simulations. Results were qualitatively similar to those observed in the primary set of simulations, with a few minor differences.

5 | CASE STUDY

We provide a brief case study to illustrate the utility of matching with replacement. The case study examines a setting in which the majority of subjects are treated or exposed, and thus in which conventional matching without replacement would be potentially problematic.

5.1 | Data sources and methods

The data for the current case study consisted of 2342 patients discharged alive following hospitalization with an acute myocardial infarction (AMI) in Ontario, Canada between 1 April 1999 and 31 March 2001, who were current smokers at the time of hospital admission, and for whom there was evidence of either smoking cessation counselling or

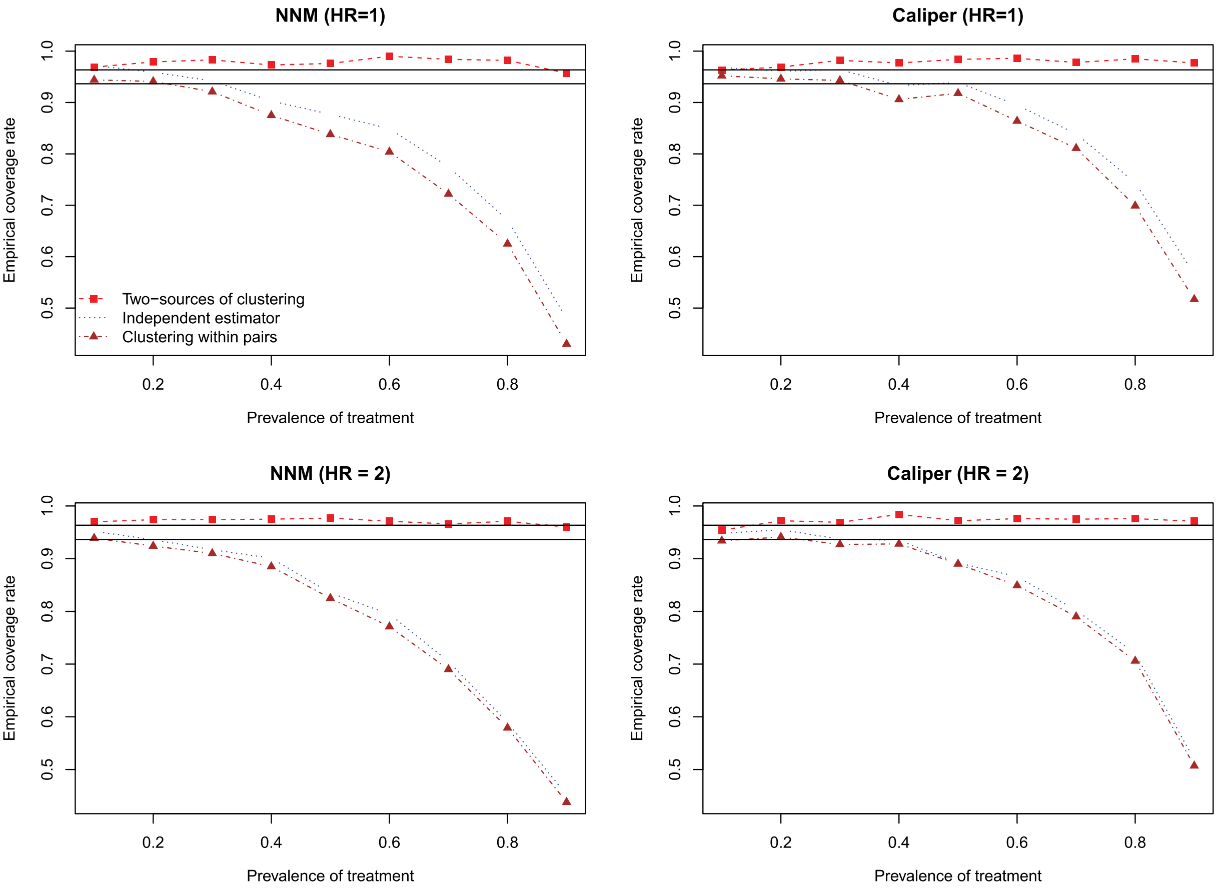
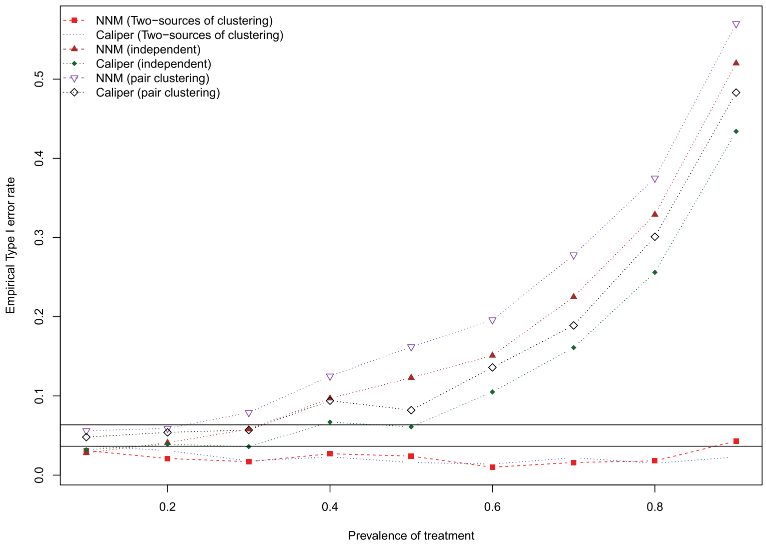


FIGURE 8 Coverage rates of estimated confidence intervals (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 9 Empirical type I error rate [Colour figure can be viewed at wileyonlinelibrary.com]



of the absence of such counselling in the patient's medical records. These data form a subset of the data which were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, a study to improve the quality of care for patients with cardiovascular disease in Ontario.[36] These data were used previously in an extensive tutorial on the application of propensity score methods.[37] However, matching with replacement was not examined in that tutorial.

For the purposes of the current case study, the treatment or exposure of interest was whether the patient received in-patient smoking cessation counseling. Of the 2342 subjects in the sample, 1588 (67.8%) received in-patient smoking

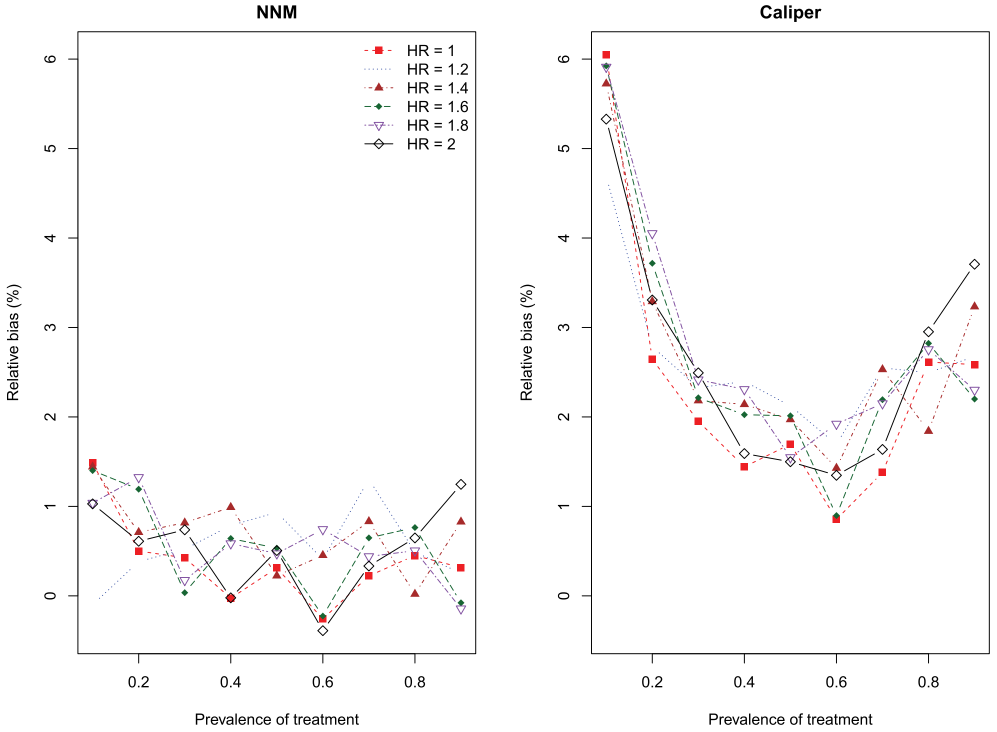


FIGURE 10 Relative bias in estimated hazard ratio [Colour figure can be viewed at wileyonlinelibrary.com]

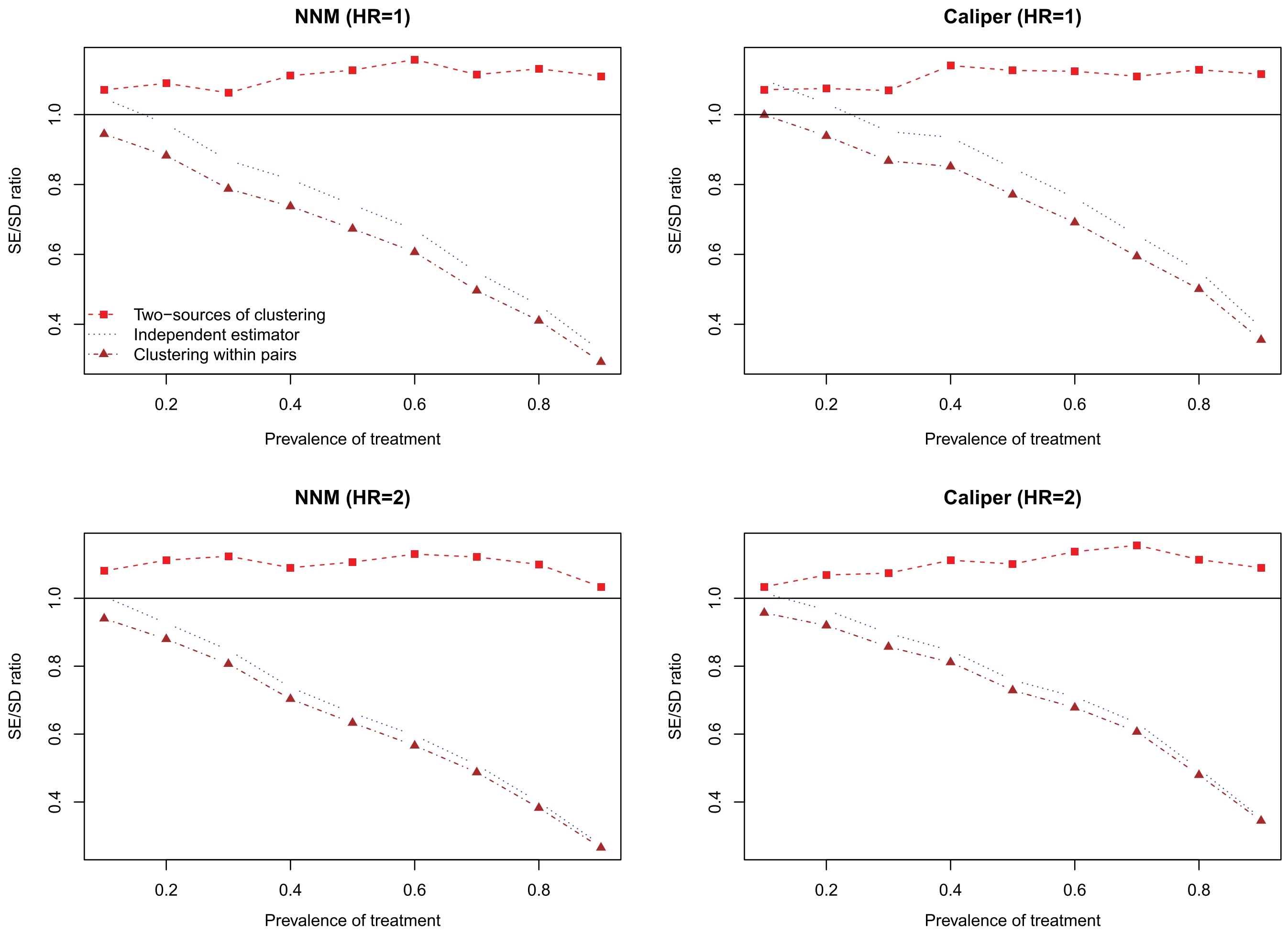


FIGURE 11 Ratio of mean estimated standard error to standard deviation of estimated log-hazard ratio (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

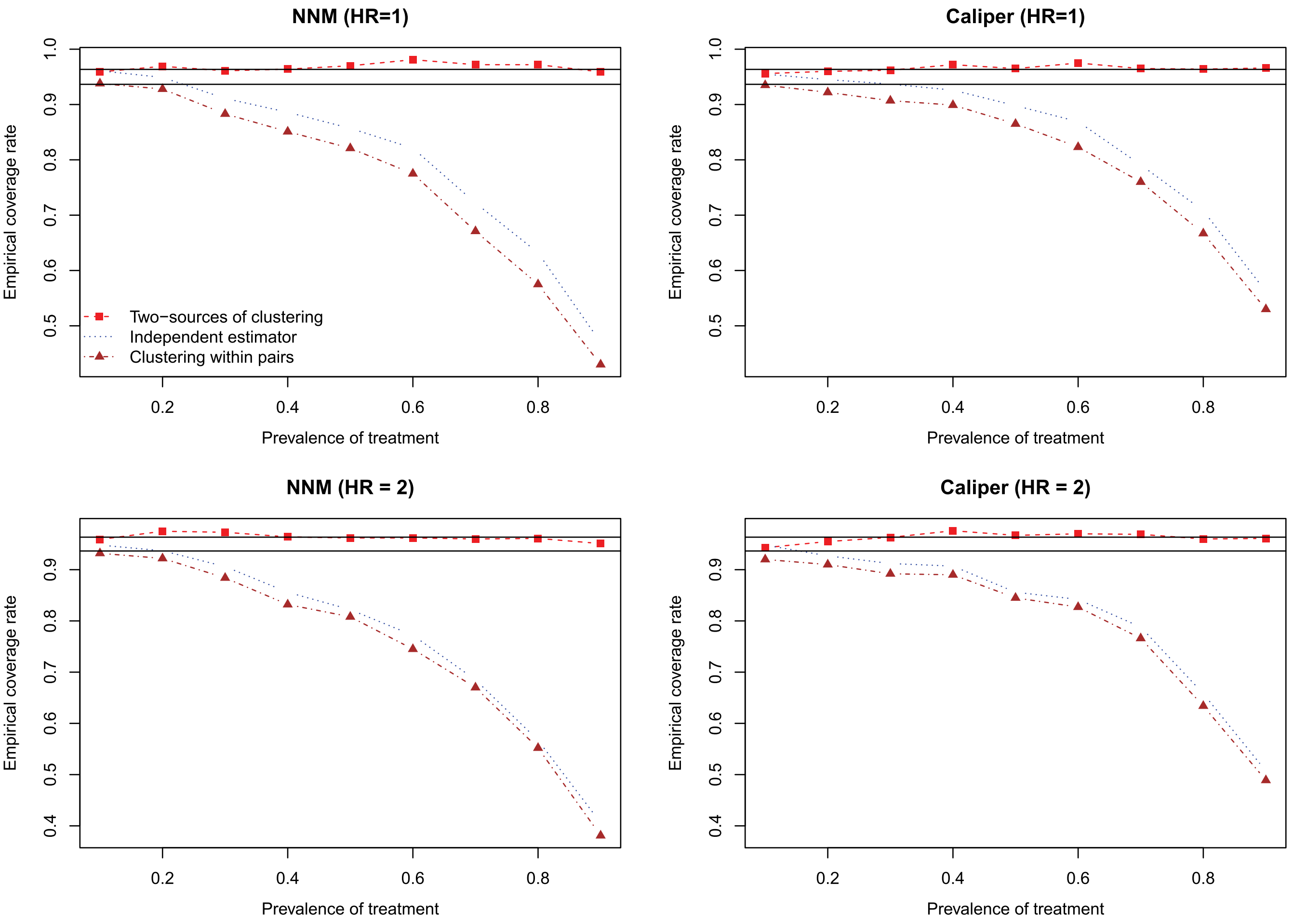


FIGURE 12 Coverage rates of estimated confidence intervals (HR = 1 and 2) [Colour figure can be viewed at wileyonlinelibrary.com]

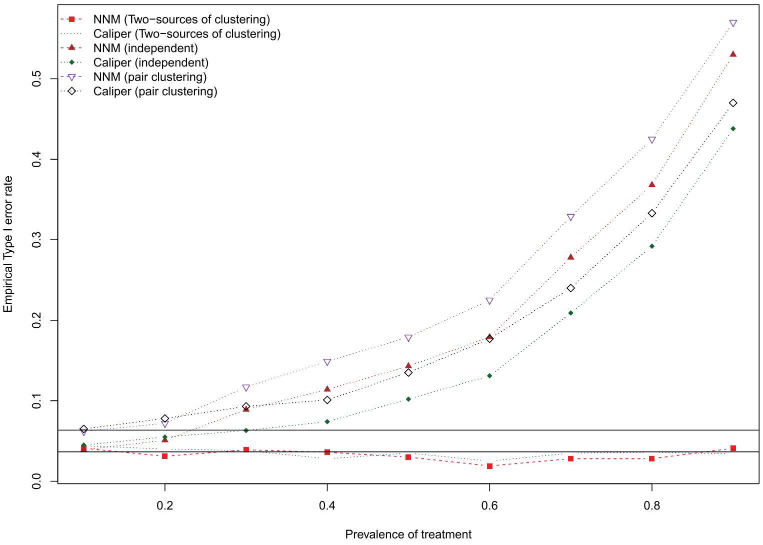


FIGURE 13 Empirical type I error rate [Colour figure can be viewed at wileyonlinelibrary.com]

cessation counseling and 754 (32.2%) did not. Thus, the majority of subjects were treated, and conventional matching without replacement would not be expected to perform well.

Patient records were linked to the Registered Persons Database using encoded versions of the patient's health card numbers, which allowed for determining the vital status of each patient. For the current study, the outcome was time to death, with subjects censored after 3 years if they were alive at that time.

A propensity score for in-patient smoking cessation counselling was estimated by using logistic regression to regress the receipt of smoking cessation counselling on 33 variables that were ascertained from the patients' medical records. These included demographic characteristics (age and sex), presenting signs and symptoms (acute pulmonary edema), vital signs on admission (systolic blood pressure, diastolic blood pressure, heart rate, and respiratory rate), classic cardiac risk factors (diabetes, hyperlipidemia, hypertension, and family history of coronary artery disease), comorbid conditions and vascular history (cerebrovascular accident/transient ischemic attack, angina, cancer, dementia, previous myocardial infarction, asthma, depression, peptic ulcer disease, peripheral vascular disease, previous coronary revascularization, and chronic congestive heart failure), laboratory test (glucose, white blood count, hemoglobin, sodium, potassium, and creatinine), and prescriptions for cardiovascular medications at hospital discharge (statin, beta-blocker, ACE inhibitor/Angiotensin receptor blockers, plavix, and ASA). Treated subjects were matched to control subjects on the logit of the propensity score using two different methods: (a) NNM with replacement and (b) NNM with a caliper equal to 0.2 of the SD of the logit of the propensity score. In each of the two matched samples, the hazard of mortality was regressed on an indicator variable denoting treatment status. The variance estimator described in Section 2 was used to estimate the SE of the log-hazard ratio. We did not consider the two alternative variance estimators that were examined in the simulations, as they were shown to have inferior performance to the proposed variance estimator.

For comparative purposes, inverse probability of treatment weights for estimating the ATT was computed. These weights are defined as $Z + \frac{e}{1-e}(1 - Z)$, where e denotes the estimated propensity score and Z is an indicator variable denoting treatment status. The hazard of the outcome was regressed on an indicator variable denoting treatment status. A 95% confidence interval was constructed using a percentile-based bootstrap confidence interval with 2000 bootstrap samples.[30,38] The inclusion of this approach allows us to compare the estimate obtained using matching with replacement with that obtained using an established estimator with the same target estimand (the ATT).

5.2 | Results

Prior to matching, the absolute standardized differences for the 33 baseline covariates ranged from 0 to 0.39, with a median of 0.16 (25th and 75th percentiles: 0.07 and 0.22).[39] Thus, there was strong evidence of systematic differences between smokers offered smoking cessation counselling and smokers not offered smoking cessation counselling.

When using NNM, all 1588 treated subjects were matched to a control subject. Of the 754 control subjects, 526 were matched to at least one treated subject, while 228 were not matched to any treated subject. Of the 526 matched control subjects, the number of treated subjects to which a control was matched ranged from 1 to 21, with a median of 2 (25th and 75th percentiles: 1 and 4, respectively). The absolute standardized differences for the 33 baseline covariates after matching ranged from 0 to 0.09, with a median of 0.02 (25th and 75th percentiles: 0.01 and 0.05). Matching on the propensity score removed the systematic differences between treated and control subjects. The estimated hazard ratio for smoking cessation counselling in the matched sample was 0.815 and the associated 95% confidence interval was (0.605, 1.097). As the confidence interval includes the null value, the estimated hazard ratio is not statistically significantly different from the null.

When using caliper matching, all 1588 treated subjects were matched to a control subject. Of the 754 control subjects, 573 were matched to at least one treated subject, while 181 were not matched to any treated subject. Of the 573 matched control subjects, the number of treated subjects to which a control was matched ranged from 1 to 10, with a median of 2 (25th and 75th percentiles: 1 and 4, respectively). The absolute standardized differences for the 33 baseline covariates after matching ranged from 0 to 0.06, with a median of 0.02 (25th and 75th percentiles: 0.01 and 0.03). Matching on the propensity score removed the systematic differences between treated and control subjects. The estimated hazard ratio for smoking cessation counselling in the matched sample was 0.989 and the associated 95% confidence interval was (0.729, 1.341). As the confidence interval includes the null value, the estimated hazard ratio is not statistically significantly different from the null.

In the sampling weighted using the inverse probability of treatment weights, the absolute weighted standardized differences ranged from 0 to 0.05, with a median of 0.01 (25th and 75th percentiles: 0.01 and 0.03). When using inverse probability of treatment weighting with ATT weights, the estimated hazard ratio was 0.849, with associated 95% confidence interval (0.677, 1.078).

6 | DISCUSSION

We conducted an extensive series of Monte Carlo simulations to examine the performance of matching with replacement on the propensity score to estimate marginal hazard ratios. We found that matching with replacement on the propensity score permitted essentially unbiased estimation of the underlying marginal hazard ratio. The proposed variance estimator resulted in estimated SEs that overestimated the sampling variation of the log-hazard ratio. However, in most scenarios, the relative bias in the estimated SEs was less than 30%. As a consequence, estimated confidence intervals and tests of the null-hypothesis were conservative, with empirical coverage rates exceeding the advertised rates and empirical type I error rates lower than the advertised rates.

The magnitude of overestimation of the standard error increased as the proportion of controls that were recycled increased (compare Figures 1 and 3). We hypothesize that the overestimation was due, in part, to a failure to fully account for the within-subject homogeneity in outcomes when constructing the second component of the variance estimator (V_2). The second component of the variance estimator accounts for within-subject correlation in outcomes. However, in our application, this correlation is by necessity one, whereas in most other applications of the robust variance estimator, this correlation will be less than one. This problem would be exacerbated with increasing recycling of controls because more of the data would exhibit this within-person correlation.

We compared the performance of our proposed variance estimator with two alternative variance estimators. While the proposed variance estimator tended to result in estimated standard errors that modestly overestimated the sampling variability of the log-hazard ratio, the two alternative estimators tended to underestimate the sampling variability of the log-hazard ratio. Furthermore, the degree of underestimation was substantial when the prevalence of treatment was high, which is particularly those settings in which matching with replacement would be an attractive strategy. Similarly, estimated confidence intervals constructed using the proposed variance estimator were conservative, while those constructed using the alternative variance estimators tended to be anticonservative. These findings suggest that the proposed variance estimator can be used with matching with replacement when estimating hazard ratios until an improved variance estimator is developed.

Propensity score matching is frequently used in the medical and epidemiological literature. However, propensity score matching requires a pool of potential controls that is larger than the number of treated subjects. Failure to identify a match for all treated subjects can result in bias due to incomplete matching.[6] As a result, matching without replacement on the propensity score may not perform well when treatments or exposures are common. In settings such as these, matching with replacement could be an attractive alternative. However, the utility of matching with replacement on the propensity score is limited by the absence of variance estimators for measures of effect that are common in the medical and epidemiological literature. While Hill and Reiter[21] have described a variety of variance estimators for use with differences in means of continuous outcomes (both for matching with and without replacement), no corresponding estimators have been described for use with matching with replacement when estimating relative risks, risk differences, or hazard ratios. These measures of effect are common in the medical and epidemiological literature. Hill and Reiter primarily evaluated the performance of different variance estimators by examining the empirical coverage rates of 95% confidence intervals for differences in means. They did not formally compare the estimated standard errors to the SD of the estimated regression coefficients. Similar to the current study, they found that, in general, the coverage rates of confidence intervals were conservative, with empirical coverage rates that tended to exceed the advertised rate. Given that Abadie and Imbens suggest that the use of the bootstrap to estimate standard errors is inappropriate when using matching with replacement,[22] it is important to develop variance estimators for use with matching with replacement when estimating common measures of effect such as the hazard ratio. We would highlight that we have not formally developed a new variance based on the sampling distribution of the log-hazard ratio when using matching with replacement. Instead, we modified a previously described variance estimator for use with generalized linear models with clustered data and have adapted it for use with the Cox model when using matching with replacement. We found that this proposed variance estimator performs relatively well.

There are certain limitations to the current study. Our findings were based on an extensive series of Monte Carlo simulations in which we considered 54 different scenarios. As such, our findings warrant replication in different scenarios and under different assumptions about the distribution of baseline covariates and about the number of measured baseline covariates and their relationship with treatment-selection and with outcome. Given our focus on estimating marginal hazard ratios, analytic determination of the performance of the variance estimator would be difficult. Furthermore, we would note that several prior studies examining the performance of propensity score methods for estimating treatment effects have employed Monte Carlo simulations.[17,23,40-44]

We conducted two additional sets of Monte Carlo simulations. First, we examined the consequences of including all measured baseline variables in the propensity score model. We observed results that were qualitatively similar to those from the primary analysis. Thus, in some settings, there may not be a meaningful decay in performance due to the inclusion of instrument variables in the propensity score model. These simulations were designed to address the performance of this estimator in settings in which the analyst may not know the true association of each baseline variable with exposure status and with the outcome. Second, we examined the consequence of misspecifying the propensity score model. In particular, we considered the consequence of omitting two interactions and a quadratic term from the propensity score model. We found that estimation of hazard ratios, standard errors, and confidence intervals tended to be robust to model misspecification. A limitation to this set of simulations is that, due to space constraints, we were not able to examine a wider range of misspecifications. However, our results suggest that estimation may be robust to modest degrees of model misspecification.

We provide the following recommendations when using matching with replacement and estimating marginal hazard ratios. First, when the prevalence of treatment is relatively low (≤30%), then a robust variance estimator that accounts for clustering within pairs can be used, as this tended to result in the most accurate estimates of the sampling variability of the log-hazard ratio. Second, when the prevalence of treatment is high (eg, ≥50%), then the proposed variance estimator should be preferred. We note that the first setting (of low treatment prevalence) is likely to be a setting in which conventional pair-matching without replacement is likely to perform well, and matching without replacement may not be necessary.

In summary, we have proposed a variance estimator for the log-hazard ratio when matching on the propensity score with replacement. Use of this variance estimator resulted in confidence intervals and statistical hypothesis tests that were modestly conservative. The use of matching with replacement will be of use in settings with common exposures or treatments. It will also be of use in settings in which matching without replacement leads to an unacceptable number of unmatched treated subjects due to an insufficient number of control subjects in the region of the distribution of the propensity score in which many treated subjects lie.

ACKNOWLEDGEMENTS

This study was supported by 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 Mid-Career Investigator award from the Heart and Stroke Foundation of Ontario. The data sets used for this study were held securely in a linked, deidentified form, and analyzed at ICES. While data sharing agreements prohibit ICES from making the data set publicly available, access may be granted to those who meet prespecified criteria for confidential access, available at www.ices.on.ca/DAS.

ORCID

Peter C. Austin https://orcid.org/0000-0003-3337-233X
Guy Cafri https://orcid.org/0000-0002-1743-429X

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. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46:399-424.
3. Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg*. 2007;134(5):1128-1135.

4. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037-2049.
5. Austin PC. A report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review and suggestions for improvement. *Circ Cardiovasc Qual Outcomes*. 2008;1:62-67.
6. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38.
7. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat*. 1993;2:405-420.
8. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*. 2000;56(1):118-124.
9. Rosenbaum PR. A characterization of optimal designs for observational studies. *J Royal Stat Soc Series B*. 1991;53:597-610.
10. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51(1):171-184.
11. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33(6):1057-1069.
12. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86:4-29.
13. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press; 2007.
14. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13(4):279-313.
15. Austin PC, Type I. Error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *Int J Biostat*. 2009;5(1). https://doi.org/10.2202/1557-4679.1146.
16. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Stat Med*. 2011;30(11):1292-1301.
17. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32(16):2837-2849.
18. Gayat E, Resche-Rigon M, Mary JY, Porcher R. Propensity score applied to survival data analysis through proportional hazards models: a Monte Carlo study. *Pharm Stat*. 2012;11(3):222-229.
19. Abadie AI, G.W. I. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006;74(1):235-267.
20. Lin DY, Wei LJ. The robust inference for the proportional hazards model. *J Am Stat Assoc*. 1989;84(408):1074-1078.
21. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med*. 2006;25(13):2230-2256.
22. Abadie A, Imbens GW. Notes and comments on the failure of the bootstrap for matching estimators. *Econometrica*. 2008;76(6): 1537-1557.
23. 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. *Stat Med*. 2007;26(4):754-768.
24. Austin PC, Stuart EA. Optimal full matching for survival outcomes: a method that merits more widespread use. *Stat Med*. 2015;34(3949):3967.
25. Austin PC, Stuart EA. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Stat Methods Med Res*. 2017;26(4):1654-1670.
26. Miglioretti DL, Heagerty PJ. Marginal modeling of nonnested multilevel data using standard software. *Am J Epidemiol*. 2007;165(4):453-463.
27. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics*. 2001;57(1):126-134.
28. Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. *Stat Med*. 2005;24(11):1713-1723.
29. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat Med*. 2014;33(24):4306-4319.
30. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655.
31. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Stat Methods Med Res*. 2016;25(5):2214-2237.
32. Austin PC, Thomas N, Rubin DB. Covariate-adjusted survival analyses in propensity-score matched samples: imputing potential time-to-event outcomes. *Stat Methods Med Res*. 2020;29:728-751. https://doi.org/10.1177/0962280218817926.
33. 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. *Stat Med*. 2007;26(4):734-753.
34. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011;10:150-161.
35. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
36. Tu JV, Donovan LR, Lee DS, et al. Effectiveness of public report cards for improving the quality of cardiac care: the EFFECT study: a randomized trial. *Jama*. 2009;302(21):2330-2337.
37. 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. *Multivar Behav Res*. 2011;46:119-151.

38. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1993.
39. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med*. 2009;28(25):3083-3107.
40. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med*. 2010;29(3):337-346.
41. Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol Drug Saf* . 2008;17(6):546-555.
42. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007;26(16): 3078-3094.
43. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol*. 2008;61(6):537-545.
44. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137-2148.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Austin PC, Cafri G. Variance estimation when using propensity-score matching with replacement with survival or time-to-event outcomes. *Statistics in Medicine*. 2020;39:1623–1640. https://doi.org/10.1002/sim.8502