The International Journal of Biostatistics

Volume 5, *Issue* 1 2009 *Article* 13

Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses

Peter C. Austin[*]

[*]Institute for Clinical Evaluative Sciences, peter.austin@ices.on.ca

Type I Error Rates, Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses*

Peter C. Austin

Abstract

Propensity-score matching is frequently used in the medical literature to reduce or eliminate the effect of treatment selection bias when estimating the effect of treatments or exposures on outcomes using observational data. In propensity-score matching, pairs of treated and untreated subjects with similar propensity scores are formed. Recent systematic reviews of the use of propensity-score matching found that the large majority of researchers ignore the matched nature of the propensity-score matched sample when estimating the statistical significance of the treatment effect. We conducted a series of Monte Carlo simulations to examine the impact of ignoring the matched nature of the propensity-score matched sample on Type I error rates, coverage of confidence intervals, and variance estimation of the treatment effect. We examined estimating differences in means, relative risks, odds ratios, rate ratios from Poisson models, and hazard ratios from Cox regression models. We demonstrated that accounting for the matched nature of the propensity-score matched sample tended to result in type I error rates that were closer to the advertised level compared to when matching was not incorporated into the analyses. Similarly, accounting for the matched nature of the sample tended to result in confidence intervals with coverage rates that were closer to the nominal level, compared to when matching was not taken into account. Finally, accounting for the matched nature of the sample resulted in estimates of standard error that more closely reflected the sampling variability of the treatment effect compared to when matching was not taken into account.

KEYWORDS: propensity score, matching, propensity-score matching, variance estimation, coverage, simulations, type I error, observational studies

*The Institute for Clinical Evaluative Sciences (ICES) is supported in part by a grant from the Ontario Ministry of Health and Long Term Care. The opinions, results and conclusions are those of the authors and no endorsement by the Ministry of Health and Long-Term Care or by the Institute for Clinical Evaluative Sciences is intended or should be inferred. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR) (MOP 86508). Dr. Austin is supported in part by a Career Investigator award from the Heart and Stroke Foundation of Ontario. The EFFECT study was funded by a Canadian Institutes of Health Research (CIHR) Team Grant in Cardiovascular Outcomes Research. The study funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

1. Introduction

The propensity score is defined as a subject's probability of exposure to a specific treatment conditional on observed baseline covariates (Rosenbaum and Rubin 1983; Rosenbaum and Rubin 1984; Austin et al. 2005). Rosenbaum and Rubin (1983) demonstrated that conditional on the propensity score, the distribution of measured independent baseline covariates is independent of treatment assignment. Therefore, treated and untreated subjects with the same propensity score will have a similar distribution of baseline covariates. Rosenbaum and Rubin (1983) further demonstrated that conditioning on the propensity score allows for unbiased estimation of linear treatment effects. Similar results have been shown for estimating rate ratios (Austin et al. 2007b) and relative risks when subject-specific relative risks are uniform (Austin 2008a). Rosenbaum and Rubin's initial article on the propensity score described three methods to estimate effects of treatments or exposures: stratification (subclassification), matching, and covariate adjustment (1983). In propensity-score matching, matched sets of treated and untreated subjects with similar propensity scores are formed. The most common implementation of propensity-score matching is pair matching, in which pairs of treated and untreated subjects are formed. In subsequent articles, Rosenbaum and Rubin examined methods that incorporate the propensity score when matching treated and untreated subjects (1985a) and the bias that can arise from incomplete or inexact matching (1985b).

Propensity-score matching is frequently used in the medical literature to reduce or eliminate the effect of treatment-selection bias, when estimating the effects of treatments and exposures on outcomes using non-randomized data. A recent systematic review of propensity-score matching in the medical literature between 1996 and 2003 found that the majority of studies did not account for the matched nature of the propensity-score matched sample when estimating the significance of the treatment effect (Austin 2008b). Similar findings were observed in a comparable systematic review of propensity score matching in both the cardiovascular surgery literature and the general cardiology literature of a more recent era (Austin 2007a; Austin 2008c). However, matched treated and untreated subjects have similar propensity scores, and thus have baseline covariates that come from the same multivariate distribution. Consequently, matched subjects are, on average, more similar in baseline covariates than are two randomly selected treated and untreated subjects; therefore, by construction, the propensity-score matched sample does not consist of independent observations. Furthermore, in the presence of confounding, baseline covariates are associated with both treatment selection and outcomes; therefore, matched treated and untreated subjects are more likely to have similar outcomes than are randomly selected treated and untreated subjects.

The objective of the current study was to examine the impact on Type I error rates, coverage of confidence intervals, and variance estimation when the matched nature of the propensity-score matched sample was not taken into account. We examined estimation of difference in means, odds ratios, hazard ratios, rate ratios, and relative risks. Monte Carlo simulations were used to determine the impact on statistical inference. In Section 2, we examine the impact on Type I error rates when the matched nature of the sample was not taken into account. In Section 3, we examine coverage of confidence intervals and variance estimation in the presence of a non-null treatment effect when the matched nature of the sample was not taken into account. In Section 4, we consider an empirical case study, in which we illustrate that differing inferences can be obtained depending on whether one accounts for the matched nature of the sample. In Section 5, we summarize our findings.

2. Type I error rates

2.1. Monte Carlo simulations - Methods

In this section, we describe the Monte Carlo simulations used to examine the impact on type I error rates when the matched nature of the propensity-score matched sample was not accounted for in the statistical analyses. We conducted separate Monte Carlo simulations for examining inferences on differences in means, odds ratios, hazard ratios, rate ratios, and relative risks. These are described separately in the following sub-sections.

2.1.1. Differences in means

We randomly generated 9 baseline covariates for each of 10,000 subjects. We used a design similar to those used in prior studies (Austin et al. 2007b; Austin 2008a; Austin 2007b; Austin et al. 2007a). We assumed that there were 9 covariates: 6 related to treatment selection and 6 related to outcome, as described in the following grid:

	Strongly associated with treatment selection	Moderately associated with treatment selection	Independent of treatment selection
Strongly associated with outcome	x_1	x_2	x_3

Moderately associated with outcome	x_4	x_5	x_6
Independent of outcome	x_7	x_8	x_9

Each covariate was randomly generated from independent standard normal distributions. We assumed the following model:

$$\text{logit}\,(p_{treat,i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_4 x_{4,i} + \beta_5 x_{5,i} + \beta_7 x_{7,i} + \beta_8 x_{8,i} \qquad (1)$$

where the logit of the probability of treatment selection for the ith subjects, $p_{treat,i}$, is linearly related to the six covariates associated with treatment selection. In the data-generating process, we assumed $\beta_1 = \beta_4 = \beta_7 = \log(2)$ and $\beta_2 = \beta_5 = \beta_8 = \log(1.5)$. We set β_0=0, so that approximately 50% of subjects would be exposed to the treatment. We then generated a treatment status for each subject (T_i) from a Bernoulli distribution with subject-specific parameter $p_{treat,i}$. For each subject, we then generated an outcome from the following model:

$$y_i = \alpha_0 + \alpha_T T_i + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} + \varepsilon_i \qquad (2)$$

where $\varepsilon_i \sim N(0,\sigma^2)$. We set $\alpha_0 = 0$, $\alpha_1 = \alpha_2 = \alpha_3 = 2$ and $\alpha_4 = \alpha_5 = \alpha_6 = 1$. Since we are examining the Type I error rate, we generated data under the null hypothesis of no treatment effect (α_T= 0). Therefore, data were generated such that, on average, treatment exposure did not have an independent effect on the outcome variable. The variance of the subject-specific error term (σ^2) was chosen so that the 6 variables related to the outcome ($x_1 - x_6$) explained a specified proportion of the variation in the outcome. We examined 5 different values of R^2: 0.02, 0.05, 0.10, 0.25, and 0.50. Thus, variation in the 6 covariates explained 2%, 5%, 10%, 25%, and 50% of the variation in the continuous response variable.

Once data had been randomly generated for each of 10,000 subjects, we estimated the propensity score model using a logistic regression model containing the six variables associated with treatment selection $(x_1, x_2, x_4, x_5, x_7, x_8)$. Treated and untreated subjects were then matched on the logit of the propensity-score using calipers of width 0.2 standard deviations of the logit of the propensity score (Austin et al. 2007a; Austin and Mamdani 2006). The difference in mean outcome between matched treated and untreated subjects was compared in two methods. First, a paired t-test was used; this method accounted for the matched-pairs design. Second, a standard t-test was used; this method did not account for the matched nature of the design. The standard t-test did not assume equal variances in the two treatment groups. The statistical significance of each test was noted. The treatment effect was classified as statistically different from zero if the p-value of the associated test was less than 0.05.

The above process was repeated 7,300 times. The proportion of simulations in which the null hypothesis of no treatment effect was rejected was determined across the 7,300 simulated datasets. The use of 7,300 simulated datasets allowed us to detect a type I error of greater than 0.055 or less than 0.045 as significantly different than 0.05, using a standard test based on the normal approximation to the binomial distribution (Rosner 1995). The above process was repeated for R^2 of 0.02, 0.05, 0.10, 0.25, and 0.50.

2.1.2. Odds ratios

A prior systematic review found that propensity score methods were most frequently used in the medical literature to estimate odds ratios and hazard ratios (Sturmer et al. 2006). We modified the data-generating process described in 2.1.1 to generate dichotomous outcomes. We modified formula (2) as follows:

$$\text{logit}(p_{outcome,i}) = \alpha_0 + \alpha_T T_i + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} \qquad (3)$$

where $p_{outcome,i}$ is the probability of the outcome for the i^{th} subject. We then generated dichotomous outcomes for each subject from a Bernoulli distribution with subject-specific parameter $p_{outcome,i}$. In the data-generating process, we set $\alpha_0 = -1.0986$, $\alpha_1 = \alpha_2 = \alpha_3 = \log(2)$ and $\alpha_4 = \alpha_5 = \alpha_6 = \log(1.5)$. We randomly generated data from 10,000 subjects. Propensity-score matching was implemented as described in Section 2.1.1. We then estimated the treatment effect in the propensity-score matched sample on the odds ratio scale. This was done in two different ways. First, we used a conventional logistic regression model estimated using maximum likelihood estimation. This model regressed the binary outcome on a dichotomous variable indicating treatment status. Model-based standard errors were used to determine the statistical significance of the treatment effect. Second, we used a logistic regression model estimated using Generalized Estimating Equation (GEE) methods (Diggle et al. 1994). This model accounted for the matched-nature of the sample. As above, 7,300 randomly generated datasets were constructed.

2.1.3. Hazard ratios

We modified the data-generating process described in 2.1.1 to generate time-to-event outcomes. We used a data-generating process described elsewhere (Bender et al. 2005), and that we have used in a prior study (Austin et al. 2007b). We modified formula (2) as follows:

$$\lambda = \alpha_T T_i + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} \tag{4}$$

We set $\alpha_1 = \alpha_2 = \alpha_3 = \log(2)$ and $\alpha_4 = \alpha_5 = \alpha_6 = \log(1.5)$. We then randomly generated time-to-event outcomes from the following formula:

$$\left(\frac{-\log(U)}{0.000001 e^{\lambda}} \right)^{1/2} \tag{5}$$

where U is a random variable from a standard uniform distribution. We designed the data-generating process so that there was no censoring. Propensity-score matching was done as described in Section 2.1.1. We then estimated the treatment effect and its statistical significance in the propensity-score matched sample using two different methods. First, we fit a univariate Cox proportional hazards regression model (Cox and Oakes 1984), which regressed survival time on a dichotomous variable denoting exposure status. This method did not take into account the matched nature of the propensity-score matched sample. Model-based standard errors were obtained. Second, we fit a Cox proportional hazards model that stratified on matched pairs (Therneau and Grambsch 2000). This method accounted for the matched nature of the propensity-score matched sample.

2.1.4. Rate ratios

We modified the data-generating process described in 2.1.1 to generate count outcomes. We modified formula (2) as follows:

$$\log(\eta_i) = \alpha_0 + \alpha_T T_i + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} \tag{6}$$

where η_i is the mean outcome for the i$^{\text{th}}$ subject. We then generated a count outcome for each subject from a Poisson distribution with subject-specific parameter η_i. In the data-generating process, we set $\alpha_0 = 0$,

$\alpha_1 = \alpha_2 = \alpha_3 = \log(2)$ and $\alpha_4 = \alpha_5 = \alpha_6 = \log(1.5)$. We randomly generated data for 10,000 subjects. Propensity-score matching was done as described in Section 2.1.1. We then estimated the treatment effect in the propensity-score matched sample on the log-rate ratio scale. This was done in two different ways. First, we used a conventional Poisson regression model estimated using maximum likelihood estimation. This model regressed the count outcome on a dichotomous variable indicating treatment status. Model-based standard errors were used to determine the statistical significance of the treatment effect. Second, we used a logistic regression model estimated using Generalized Estimating Equation (GEE) methods (Diggle et al. 1994). This model accounted for the matched-nature of the sample. As above, 7,300 randomly generated datasets were constructed.

2.1.5. Relative risks

We modified the data-generating process described in 2.1.1 to generate dichotomous outcomes. However, unlike in Section 2.1.2, we estimated the treatment effect on the relative risk scale, rather than the odds ratio scale. In this section, the 9 covariates ($x_1 - x_9$) were generated from independent Bernoulli distributions, each with a parameter of 0.5. We then randomly generated an exposure status for each subject, as described in Section 2.1.1. However, we set $\beta_0 = -3.5$, as this resulted in approximately 50% of the subjects being exposed to the treatment when the covariates were binary. We modified formula (2) as follows:

$$\log(p_{outcome,i}) = \alpha_0 + \alpha_T T_i + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \alpha_5 x_{5,i} + \alpha_6 x_{6,i} \qquad (7)$$

where $p_{outcome,i}$ is the subject-specific probability of the binary outcome. We then generated a dichotomous outcome for each subject from a Bernoulli distribution with subject-specific parameter $p_{outcome,i}$. In the data-generating process, we set $\alpha_0 = -3.5$, $\alpha_1 = \alpha_2 = \alpha_3 = \log(2)$ and $\alpha_4 = \alpha_5 = \alpha_6 = \log(1.5)$. Due to the use of the logarithmic link function, the linear predictor in formula (7) was constrained to be less than zero (so that the probability of the outcome lies between 0 and 1). Therefore, we could not use covariates from a distribution with support over the real line, as the linear predictor could have exceeded 0 for some subjects. For this reason, baseline covariates were generated from Bernoulli distributions rather than from normal distributions. We then randomly generated data for 10,000 subjects. Propensity-score matching was done as described in Section 2.1.1.

We estimated the relative risk and its statistical significance using two different methods; first, in a manner that did not account for the matched nature of the sample (Rosner 1995); second, we used methods appropriate for a matched-pairs design as described by Agresti and Min (2004). In the analyses that did not account for matching, assume that: a treated subjects experienced the outcome, b treated subjects did not experience the outcome, c untreated subjects experienced the outcome, and d untreated subjects did not experience the outcome. Then, the standard error of the log-relative risk was estimated as:

$$se[\ln(RR)] \cong \sqrt{\frac{b}{an_1} + \frac{d}{cn_0}}$$

where n_1 and n_0 are the number of treated and untreated subjects, respectively (in this setting we have $n_1 = n_0$ by design). 95% confidence intervals were constructed for the log-relative risk using normal-theory methods. In the analyses that accounted for the matched-pairs design assume that there were: a matched pairs in which both the treated and untreated subjects experienced the event, b matched pairs in which the treated subject did not experience the outcome event while the untreated subject experienced the outcome event, c matched pairs in which the treated subject experienced the outcome event while the untreated subject did not experience the outcome event, and d matched pairs in which both the treated and untreated subjects did not experience the outcome event. Then $(a+c)/(a+b)$ is the maximum likelihood estimator of the relative risk. The estimated asymptotic standard error of the log-relative risk is given by $\sqrt{(b+c)/\{(a+b)(a+c)\}}$. 95% confidence intervals for the log-relative risk were estimated using normal-theory approximations.

2.2. Monte Carlo simulations - Results

In this section, we summarize the findings of the Monte Carlo simulations conducted to examine the impact on Type I error rates when the matched nature of the propensity-score matched samples were not taken into account when estimating the variance of the treatment effect. Results are reported in Table 1.

In Table 1, we report the empirical type I error rate for each measure of effect, when both matched and unmatched analyses were performed. Due to our use of 7,300 iterations of the Monte Carlo simulations, empirical type I error rates that exceed 0.055 or that are less than 0.045, are statistically significantly different than 0.05. We also report the significance levels associated with the use of McNemar's test to test the null hypothesis that the matched and unmatched tests had equal type I error rates.

When estimating the statistical significance of differences in means, both methods resulted in type I error rates that were not significantly different from 0.05 when the 6 predictor variables explained 2%, 5%, and 10% of the variation in the response variable. However, when the 6 predictor variables explained 25% of the variation in the response variable, then using an unmatched analysis resulted in an empirical type I error rate of 0.0421, which was significantly different from 0.05. In contrast to this, the use of a matched analysis resulted in an empirical type I error rate of 0.0470, which was not significantly different from 0.05. These two type I error rates were different from each other (P < 0.0001). When the 6 predictor variables explained 50% of the variation in the response variable, then both matched and unmatched analyses resulted in type I error rates that were less than 0.05 (0.0430 and 0.0308). Furthermore, these two type I error rates were significantly different from one another (P < 0.0001). The use of an unmatched test tended to result in more conservative tests than did the use of a matched test. The matched test had an empirical type I error rate that was closer to the advertised level.

When estimating an odds ratio, the use of a matched test resulted in an empirical type I error rate of 0.0466, whereas the use of an unmatched test resulted in an empirical type I error rate of 0.0422. The type I error rate for the matched test was not significantly different from 0.05, while that of the unmatched test was significantly different from 0.05. The two type I error rates were significantly different from one another (P < 0.0001). Similar results were obtained when estimating relative risks.

When estimating hazard ratios, both methods resulted in overly conservative tests, with type I error rates that were significantly different than 0.05. The two type I error rates were significantly different from one another (P < 0.0001): the matched analysis had an empirical type I error rate of 0.0403, whereas the unmatched analysis had an empirical type I error rate of 0.0277. The type I error rate of the matched test was closer to the advertised nominal level.

When estimating rate ratios, the matched test had an empirical type I error rate of 0.0514, whereas the unmatched analysis resulted in a type I error rate of 0.4771. The type I error rate of the matched test was not significantly different from 0.05, while the type I error rate of the unmatched test was significantly different from 0.05. These two type I error rates were significantly different from another (P < 0.0001).

3. Coverage of confidence intervals and variance estimation

In Section 2 we examined the impact on the Type I error rate when the matched nature of the propensity-score matched sample was not taken into account in estimating the significance of the treatment effect. Data were generated such that the null hypothesis was true: the treatment had no independent effect on outcomes. In the current section, we examine the coverage of confidence intervals and variance estimation when the null hypothesis is false: in the presence of a non-null treatment effect. In this section, we restrict our attention to three measures of treatment effect: differences in means, rate ratios, and relative risks. We do not consider odds ratios and hazard ratios, since prior research has demonstrated that propensity score methods result in biased estimation of odds ratios and hazard ratios and that confidence intervals do not have desired coverage rates (Austin et al. 2007b; Austin 2007b). Reasons for this are described elsewhere (Austin et al. 2007b; Austin 2007b).

3.1. Monte Carlo simulations – Methods

In this section, we describe the Monte Carlo simulations that were conducted to examine coverage of confidence intervals and variance estimation in propensity-score matched samples. These Monte Carlo simulations are variations of those described in Section 2. As above, using each data-generating process, we randomly generated 7,300 datasets. This will allow us to classify any confidence interval whose empirical coverage rate is greater than 0.955 or less than 0.945 as having a coverage rate that is significantly different than 0.95.

3.1.1 Difference in means

These Monte Carlo simulations were similar in design to those described in Section 2.1.1; however, we considered two different non-null treatment effects: α_T = 1 and 2 (see formula (2)). Propensity-score matching was done as in the simulations described in Section 2.1.1. Within each propensity-score matched sample we first estimated the treatment effect as the difference in means between treated and untreated subjects. Second, we estimated the standard deviation of the difference in means assuming two independent samples (i.e. ignoring the matched nature of the propensity-score matched sample). The standard error of the treatment effect (the difference in means) was estimated by $\sqrt{s_1^2/n_1 + s_0^2/n_0}$, where s_1^2 and s_0^2 denote the sample variance of the outcome in the treated and untreated subjects, respectively, in the propensity-score matched sample. Similarly, n_1 and n_0 denote the number of treated and untreated subjects, respectively, in the propensity-score matched sample (in this setting, we have $n_1 = n_0$ by design). We then estimated a 95% confidence interval using Satterthwaite's Method to determine the approximate degrees of freedom for the t-distribution (Rosner 1995). Third, we estimated the standard deviation of the difference in means using methods that accounted for the matched nature of the propensity-score matched sample. Let d_i represent the observed difference in outcomes within the ith matched pair, while $\overline{d}$ denotes the mean difference in outcomes. Furthermore, let s_d denote the sample standard deviation of the d_i in the propensity-score matched sample. Then $s_d / \sqrt{n}$ is the standard error of the difference in means, where n denotes the number of matched pairs. 95% confidence intervals were constructed as $(\overline{d} - t_{n-1,0.975} s_d / \sqrt{n}, \overline{d} + t_{n-1,0.975} s_d / \sqrt{n})$ (Rosner 1995).

The above analyses were conducted in each of the 7,300 randomly generated datasets. We then determined the empirical coverage rate of the estimated 95% confidence intervals. This was estimated as the proportion of confidence intervals that contained the true treatment effect used in the data-generating process. We also computed the mean standard error of the estimated treatment effect across the 7,300 randomly generated datasets. This was compared with the standard deviation of the estimated treatment effect across the 7,300 randomly generated datasets.

3.1.2. Rate ratios

These Monte Carlo simulations were similar in design to those described in Section 2.1.4; however, we considered two different non-null treatment effects: α_T = log(1.5) and log(2). The methods and analyses were similar to those described in 3.1.1.; however, the treatment effect was estimated using a Poisson generalized linear model. For the unmatched analysis, the model was estimated using conventional maximum likelihood estimation. Model-based standard errors were computed along with Wald 95% confidence intervals. For the analysis that accounted for the matched nature of the propensity-score matched sample, the Poisson regression model was estimated using GEE methods. Robust estimates of standard errors were obtained and were used to construct 95% confidence intervals.

3.1.3. Relative risks

These Monte Carlo simulations were similar in design to those described in Section 2.1.5.; however, we considered two different non-null treatment effects: α_T = log(1.5) and log(2). $\beta_{0,treat}$ was set to -4.1 to constrain the linear predictor to less than 0 for all subjects, regardless of the treatment effect and the covariate values. Additionally, $\alpha_{0,outcome}$ was set to -4. The methods and analyses were, with some modifications, similar to those described in 3.1.1.

The above analyses were conducted in each of the 7,300 randomly generated datasets. We then determined the empirical coverage rate of the estimated 95% confidence intervals. This was estimated as the proportion of confidence intervals that contained the true treatment effect used in the data-generating process. We also computed the mean standard error of the estimated log-relative risk across the 7,300 randomly generated datasets. This was compared with the standard deviation of the estimated log-relative risk across the 7,300 randomly generated datasets.

3.2. Monte Carlo simulations – Results

In this section, we summarize the findings of the Monte Carlo simulations conducted to examine the impact on coverage of confidence intervals and variance estimation when the matched nature of the propensity-score matched sample was not taken into account in estimating the variance of the treatment effect. Results are reported in Tables 2 and 3.

In Table 2, we report the empirical coverage rates of 95% confidence intervals for matched and unmatched analyses. Due to our use of 7,300 iterations of the Monte Carlo simulations, confidence intervals whose empirical coverage rates either exceed 0.955 or are less than 0.945, have coverage rates were statistically significantly different from 0.95. We also report the significance level of McNemar's test used to test the null hypothesis that matched and unmatched analysis had confidence intervals with the same coverage rate. In Table 3, we report the empirical standard deviation of the treatment effect across the 7,300 simulated datasets and the mean standard error of the estimated treatment effects across the 7,300 simulated datasets. We also report the ratio of the mean standard error under a matched analysis to the empirical standard deviation of the treatment effect across the 7,300 simulated dataset, and the ratio of the mean standard error under an unmatched analysis to the empirical standard deviation of the treatment effect across the 7,300 simulated datasets.

When estimating differences in means, when the six covariates explained a low to moderate degree of the variation in the outcome, both matched and unmatched analyses tended to result in confidence intervals with approximately the advertised coverage rates. When the six covariates explained 50% of the variation in the outcome, then both methods resulted in confidence intervals with coverage rates that exceeded 0.95. However, the use of matched analyses resulted in coverage rates that were closer to the advertised level than did the use of unmatched analyses. Standard error estimates from both matched and unmatched analyses tended to over-estimate the standard deviation of the empirical distribution of the true treatment effect. However, in nine of the 10 scenarios, the over-estimation was greater when an unmatched analysis was employed compared to when a matched analysis was employed. For instance, when the true treatment effect was 1, and the six covariates explained 25% of the variation in the outcome, then the matched analysis over-estimated the standard deviation of the sampling distribution of the treatment effect by 2.77% whereas the unmatched analysis resulted in an over-estimation of 5.17%. Similarly, when the true treatment effect was 2, and the six covariates explained 25% of the variation in the outcome, then the matched analysis resulted in an over-estimation by 2.39%, whereas the unmatched analysis resulted in an over-estimation by 4.78%.

The use of an unmatched analysis in estimating rate ratios resulted in confidence intervals with substantially lower than advertised coverage rates. When the true rate ratios were 1.5 and 2, then, an unmatched analysis produced confidence intervals with empirical coverage rates of 0.4884 and 0.4656, respectively. In comparison, the use of matched analyses resulted in empirical coverage rates of 0.9575 and 0.9534, respectively. For both true rate ratios, coverage rates of confidence intervals were significantly different between the matched and unmatched analysis ($P < 0.0001$). The use of unmatched analyses resulted in estimates of the standard error of the treatment effect that were 64.42% and 67.22% lower than the empirical sampling standard deviation of the treatment effect. By contrast, the use of matched analyses resulted in estimates of the standard error of the treatment effect that were 5.42% and 4.56% greater than the empirical standard deviation of the treatment effect across the 7,300 simulated datasets.

The use of an unmatched analysis when estimating relative risks resulted in confidence intervals with higher than advertised coverage rates (0.9560 and 0.9608 for relative risks of 1.5 and 2, respectively). However, the use of a matched analysis resulted in confidence intervals with the advertised coverage rates (0.9495 and 0.9518, respectively). For both true relative risks, the coverage rate for confidence intervals derived from the matched analysis were different from those for the unmatched analysis ($P < 0.0001$). The use of unmatched analyses resulted in estimates of the standard error of the log-relative risk that were 3.51% and 3.17% greater than the standard deviation of the empirical sampling distribution of the log-relative risk, when the relative risks were 1.5 and 2, respectively. In contrast, the use of matched analyses resulted in estimates of the standard error of the log-relative risk that were 0.32% higher and 0.45% lower than the standard deviation of the empirical sampling distribution of the log-relative risk.

4. Case study

In this section, we present a brief case study to illustrate that differing inferences can be obtained depending on whether or not one accounts for the matched nature of the propensity-score matched sample. We used data on 9,081 patients who were discharged alive with an acute myocardial infarction (AMI or heart attack) from 102 hospitals in Ontario, Canada, between April 1, 1999 and March 31, 2001 and who did not die on the day of discharge. These data are similar to those reported elsewhere (Austin 2007c; Austin and Tu 2006; Austin and Mamdani 2006; Austin et al. 2006) and were collected as part of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, an initiative focused on improving the quality of care for cardiovascular disease patients in Ontario (Tu et al. 2004). Data on patient demographics, presenting signs and symptoms, classic cardiac risk factors, comorbid conditions and vascular history, vital signs on admission, and results of laboratory tests were abstracted directly from patients' medical records. The exposure of interest was whether the patient was prescribed a statin lipid-lowering agent at hospital discharge. Statins have been shown in several large randomized controlled trials, to reduce the risk of major coronary events (LaRosa et al. 1999).

We examined physician billing claims submitted to the Ontario Health Insurance Plan (OHIP), which provides universal insurance coverage to all residents of Ontario. We examined claims submitted by general practitioners/family physicians for visits during which an intermediate assessment was performed (fee code A007), with an associated diagnosis of *osteoarthritis and allied disorders* (coded as 715 using the International Classification of Disease, 9th Revision: ICD-9 coding scheme). The outcome of this case study was the number of primary care physician visits with this associated diagnosis within 3 years of hospital discharge. Patients in the EFFECT study were linked to the OHIP database using encrypted patient health card numbers.

Overall, 3,046 (33.5%) patients received a prescription for a statin at discharge, while 6,035 (66.5%) did not receive a prescription at discharge. A propensity score model derived in a previous study was used to predict receipt of a statin prescription at discharge (Austin and Mamdani, 2006). Treated and untreated subjects were matched on the logit of the propensity score using calipers of width equal to 0.2 of the standard deviation of the logit of the propensity score. This resulted in the formation of 2,361 matched pairs of patients who did and did not receive a statin prescription at discharge.

In the matched sample, the effect of a receipt of a statin prescription on the rate of primary care physician visits with an associated diagnosis of osteoarthritis and allied disorders was determined using Poisson regression. The number of primary care physician visits with an associated diagnosis of osteoarthritis and allied disorders was regressed on exposure to a statin prescription at hospital discharge. The logarithm of the number of days that the patient was at risk during the three years of follow-up was used as an offset variable. A standard Poisson regression, estimated maximum likelihood estimation, was used when the matched nature of the sample was not taken into account. To account for the matched nature of the sample, the model was estimated using generalized estimating equation (GEE) methods.

Among the 4,722 patients in the matched sample, the number of primary care visits with an associated diagnosis of osteoarthritis and allied disorders ranged from 0 to 33 during the three years of follow-up. For patients who did not receive a statin prescription at hospital discharge, the rate of primary care physician visits for the given diagnosis was 0.105 per patient per year, while for the exposed patients it was 0.091 visits per patient per year. Statin exposure at hospital discharge resulted in a 13.3% decrease in the rate of primary care physicians visits with an associated diagnosis of osteoarthritis and allied disorders (rate ratio: 0.867). When the matched nature of the sample was not taken into account, one inferred that the treatment effect was statistically significant (P = 0.0116). However, when the matched nature of the sample was taken into account, one inferred that the treatment effect was not statistically significant (P = 0.2112). Thus, using a significance level of 0.05 to denote statistical significance, one would derive differing conclusions depending on whether one accounted for the matched nature of the sample.

5. Discussion

In this paper we have examined the impact of not accounting for the matched nature of a propensity-score matched sample versus accounting for the matched design on type I error rates, coverage rates of confidence intervals, and variance estimation. We examined a wide range of measures of effect: difference in means, odds ratios, hazard ratios, rate ratios, and relative risks. In an empirical case study, we demonstrated that differing inferences can be obtained depending on whether one accounts for the matched nature of the sample.

When estimating a difference in means, we found that both matched and unmatched methods tended to have the appropriate type I error rate when the baseline covariates explained a low to moderate proportion of the variance in the outcome. However, when the baseline covariates explained a larger proportion of the variance in the outcome, a matched test had the correct type I error rate, while an unmatched test was overly conservative with a type I error rate that was less than 0.05. When estimating odds ratios, rate ratios, and relative risks, matched tests had the correct type I error rate, while unmatched tests had incorrect type I error rates. Finally, when estimating hazard rates, unmatched tests had type I error rates that were more conservative than matched tests.

When estimating non-null treatment effects, unmatched analyses tended to result in standard errors of the estimated treatment effect that overestimated the sampling variability of the treatment effect. In contrast, matched analyses resulted in estimates of the standard error of the treatment effect that were closer to the standard deviation of the sampling distribution of the treatment effect. Furthermore, for estimating rate ratios and relative risks, matched analyses resulted in confidence intervals that had coverage rates closer to the nominal level than did unmatched analyses.

A systematic review of the use of propensity-score methods in the medical literature found that they were most frequently used to analyze the effect of treatments and exposures on dichotomous or time-to-event outcomes (Sturmer 2006). Our findings suggest that in these settings, applied researchers should apply statistical methods that account for the matched nature of the propensity-score matched sample. Accounting for the matched nature of the sample will result in tests with appropriate type I error rates and confidence intervals with coverage rates that are closer to the nominal level.

In the medical literature, propensity-score methods are less frequently used to determine the effects of exposures or treatments on continuous outcomes (Sturmer 2006). Our study demonstrates that there is no advantage to employing an unmatched analysis. In contrast, a matched analysis resulted in estimates of the standard error of the treatment effect that better reflect the sampling variation of the treatment effect. Furthermore, when the baseline covariates explained a moderate proportion of the variability in the outcome, a matched analysis resulted in type I error rates and coverage rates for confidence intervals that were closer to the advertised level.

Two prior systematic reviews of propensity-score matching in the medical literature found that the large majority of published studies ignored the matched nature of the propensity-score matched sample when estimating the variance of the treatment effect (Austin 2007a 2008b). In the current study, we found that ignoring the matched nature of the propensity-score matched sample can result in tests with incorrect type I error rates, confidence intervals that do not have the advertised coverage rates, and incorrect estimates of the sampling variability of the estimates of the treatment effect. Applied researchers should employ matched analyses when estimating differences in means, odds ratios, hazard ratios, rate ratios, and relative risks.

Table 1. Empirical Type I error rates for different measures of treatment effect.

Measure of effect	Type I error rate – matched analysis	Type I error rate – unmatched analysis	P-value (McNemar's test)
Difference in means (R^2 = 0.02)	0.0485	0.0486	0.8084
Difference in means (R^2 = 0.05)	0.0486	0.0485	0.7630
Difference in means (R^2 = 0.10)	0.0485	0.0467	0.0003
Difference in means (R^2 = 0.25)	0.0470	0.0421	< 0.0001
Difference in means (R^2 = 0.50)	0.0430	0.0308	< 0.0001
Odds ratio	0.0466	0.0422	< 0.0001
Hazard ratio	0.0403	0.0277	< 0.0001
Rate ratio	0.0514	0.4771	< 0.0001
Relative risk	0.0485	0.0408	< 0.0001

Note: McNemar's test tests the null hypothesis that the Type I error rates for the two methods are equal to one another.

Table 2. Coverage of 95% confidence intervals in propensity-score matched samples.

Treatment effect	Coverage of 95% confidence interval – matched analysis	Coverage of 95% confidence interval – unmatched analysis	P-value comparing coverage rates of confidence intervals (McNemar's test)
Difference in means			
1 (R^2 = 0.02)	0.9512	0.9512	1.0000
1 (R^2 = 0.05)	0.9510	0.9511	0.7389
1 (R^2 = 0.10)	0.9432	0.9456	< 0.0001
1 (R^2 = 0.25)	0.9541	0.9585	< 0.0001
1 (R^2 = 0.50)	0.9582	0.9689	< 0.0001
2 (R^2 = 0.02)	0.9477	0.9480	0.6171
2 (R^2 = 0.05)	0.9508	0.9525	0.0047
2 (R^2 = 0.10)	0.9501	0.9527	0.0001
2 (R^2 = 0.25)	0.9490	0.9532	< 0.0001
2 (R^2 = 0.50)	0.9577	0.9669	< 0.0001
Rate ratios			
1.5	0.9575	0.4884	< 0.0001
2	0.9534	0.4656	< 0.0001
Relative risks			
1.5	0.9495	0.9560	< 0.0001
2	0.9518	0.9608	< 0.0001

Table 3. Variance estimation in propensity-score matched samples.

Treatment effect	Empirical standard deviation for treatment effect across 7,300 simulated datasets	Mean standard error of estimated treatment effect (matched analysis) across 7,300 datasets	Mean standard error of estimated treatment effect (unmatched analysis) across 7,300 datasets	Ratio of mean SE to empirical SD of treatment effect (matched analysis)	Ratio of mean SE to empirical SD of treatment effect (unmatched analysis)
Difference in means					
1 (R^2 = 0.02)	0.7203	0.7284	0.7297	1.0112	1.0131
1 (R^2 = 0.05)	0.4546	0.4581	0.4601	1.0077	1.0121
1 (R^2 = 0.10)	0.3212	0.3209	0.3237	0.9991	1.0078
1 (R^2 = 0.25)	0.1916	0.1969	0.2015	1.0277	1.0517
1 (R^2 = 0.50)	0.1242	0.1319	0.1386	1.0620	1.1159
2 (R^2 = 0.02)	0.7291	0.7283	0.7297	0.9989	1.0008
2 (R^2 = 0.05)	0.4508	0.4581	0.4601	1.0162	1.0206
2 (R^2 = 0.10)	0.3172	0.3208	0.3236	1.0113	1.0202
2 (R^2 = 0.25)	0.1923	0.1969	0.2015	1.0239	1.0478
2 (R^2 = 0.50)	0.1219	0.1318	0.1386	1.0812	1.1370
Rate ratios					
1.5	0.048	0.0506	0.0166	1.0542	0.3458
2	0.0482	0.0504	0.0158	1.0456	0.3278
Relative risks					
1.5	0.0939	0.0942	0.0972	1.0032	1.0351
2	0.0883	0.0879	0.0911	0.9955	1.0317

References

Agresti A, Min Y (2004). Effects and Non-effects of Paired Identical Observations in Comparing Proportions with Binary Matched-Pairs Data. *Statistics in Medicine*. 23:65-75.

Austin PC (2007a). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: A systematic review and suggestions for improvement. *Journal of Thoracic and Cardiovascular Surgery*. 134:1128-1135.

Austin PC (2007b). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*. 26:3078-3094.

Austin PC (2007c). A comparison of classification and regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*. 26:2937-2957.

Austin PC (2008a). The performance of different propensity score methods for estimating relative risks. *Journal of Clinical Epidemiology*. 61:537-545.

Austin PC (2008b). A critical appraisal of propensity score matching in the medical literature from 1996 to 2003. *Statistics in Medicine*. 27:2037-2049.

Austin PC (2008c). A report card on propensity-score matching in the cardiology literature from 2004 to 2006: results of a systematic review. *Circulation: Cardiovascular Quality and Outcomes*. 1:62-67.

Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV (2005). The Use of the Propensity Score for Estimating Treatment Effects: Administrative versus Clinical Data. *Statistics in Medicine*. 24:1563-1578.

Austin PC, Mamdani MM (2006). A Comparison of Propensity Score Methods: A Case-Study Estimating the Effectiveness of Post-AMI Statin Use. S*tatistics in Medicine*. 25:2084-2106.

Austin PC, Tu JV (2006). Comparing Clinical Data with Administrative Data for Producing AMI Report Cards. *Journal of the Royal Statistical Society – Series A (Statistics in Society)*. 169:115-126.

Austin PC, Mamdani MM, Juurlink DN, Alter DA, Tu JV (2006). Missed Opportunities in the Secondary Prevention of Myocardial Infarction: An

Assessment of the Effects of Statin Underprescribing on Mortality. *American Heart Journal*. 151:969-975.

Austin PC, Grootendorst P, Anderson GM (2007a). 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*. 26:734-753.

Austin PC, Grootendorst P, Normand SLT, Anderson GM (2007b). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine*. 26:754-768.

Bender R, Augustin T, Blettner M (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*. 24:1713-1723.

Cox DR, Oakes K (1984). Analysis of Survival Data. London: Chapman & Hall; 1984.

Diggle PJ, Liang KY, Zeger SL. Analysis of Longitudinal Data. Oxford: Oxford University Press.

LaRosa JC, He J, Vupputuri S (1999). Effect of statins on risk of coronary disease: A meta-analysis of randomized controlled trials. *Journal of the American Medical Association*. 282:2340-2346.

Rosenbaum PR, Rubin DB (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*. 70:41-55.

Rosenbaum PR, Rubin DB (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 79: 516-524.

Rosenbaum PR, Rubin DB (1985a). Constructing a control group by multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*. 39:33-38.

Rosenbaum PR, Rubin DB (1985b). The bias due to incomplete matching. *Biometrics*. 41:103-116.

Rosner B (1995). *Fundamentals of Biostatistics, fourth edition*. Belmont, CA: Duxbury Press.

Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*. 59:437-447.

Therneau TM, Grambsch PM (2000). Modeling Survival Data: Extending the Cox Model. New York, NY: Springer-Verlag.

Tu JV, Donovan LR, Lee DS, Austin PC, Ko DT, Wang JT, Newman AM (2004). Quality of Cardiac Care in Ontario. Institute for Clinical Evaluative Sciences: Toronto, Ontario.