Received: 19 August 2020 | Revised: 25 May 2021 | Accepted: 29 June 2021

DOI: 10.1002/sim.9141

RESEARCH ARTICLE

Statistics in Medicine WILEY

Estimating adjusted risk differences by multiply-imputing missing control binary potential outcomes following propensity score-matching

Peter C. Austin[1,2,3] | **Donald B. Rubin**[4,5,6] | **Neal Thomas**[7]

[1]ICES, Toronto, Ontario, Canada

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

[3]Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Ontario, Canada

[4]Yau Mathematical Sciences Center, Tsingua University, Beijing, China

[5]Department of Statistical Science, Fox School of Business, Temple University, Philadelphia, Pennsylvania, USA

[6]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

[7]Pfizer Global Research and Development, Pfizer, Groton, Connecticut, USA

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

Funding information
Canadian Institutes of Health Research, Grant/Award Number: PJT - 166161; Heart and Stroke Foundation of Canada, Grant/Award Number: Mid-Career Investigator Award

We describe a new method to combine propensity-score matching with regression adjustment in treatment-control studies when outcomes are binary by multiply imputing potential outcomes under control for the matched treated subjects. This enables the estimation of clinically meaningful measures of effect such as the risk difference. We used Monte Carlo simulation to explore the effect of the number of imputed potential outcomes under control for the matched treated subjects on inferences about the risk difference. We found that imputing potential outcomes under control (either single imputation or multiple imputation) resulted in a substantial reduction in bias compared with what was achieved using conventional nearest neighbor matching alone. Increasing the number of imputed potential outcomes under control resulted in more efficient estimation, with more efficient estimation of the estimated risk difference when increasing the number of the imputed potential outcomes. The greatest relative increase in efficiency was achieved by imputing five potential outcomes; once 20 outcomes under control were imputed for each matched treated subject, further improvements in efficiency were negligible. We also examined the effect of the number of these imputed potential outcomes on: (i) estimated standard errors; (ii) mean squared error; (iii) coverage of estimated confidence intervals. We illustrate the application of the method by estimating the effect on the risk of death within 1 year of prescribing beta-blockers to patients discharged from hospital with a diagnosis of heart failure.

KEYWORDS
Monte Carlo simulations, multiple imputation, propensity score, propensity score matching

1 | INTRODUCTION

Observational studies are increasingly being used to estimate the effects of treatments, that is, interventions. These studies are affected by confounding when the distribution of prognostically important covariates differs between treated and control subjects. Statistical methods must be used to minimize the effects of this confounding, so that credible inferences about treatment effects can be drawn. Methods based on estimated propensity scores are increasingly being used to reduce

the effects of measured confounding. The true propensity score is the probability of treatment assignment conditional on measured baseline covariates.[1] There are a variety of ways to use the propensity score, either alone or in combination: matching, stratification, inverse probability of treatment weighting (IPTW), and model-based covariate adjustment,[1-5] all of which have been used in the medical and epidemiological literature for decades.[6,7]

Matching on the propensity score is a popular analytic method in the medical literature.[7-9] A common approach to matching on the propensity score forms matched pairs of treated and control subjects who share a similar value of the propensity score. Two popular algorithms for matching on the propensity score are nearest neighbor matching (NNM) and NNM within specified caliper widths.[10] The latter can result in a greater reduction in bias due to the restriction on the quality of the matches.[11] However, the latter can also result in bias due to incomplete matching because a matched control subject may not be identified for all treated subjects, resulting in the exclusion of some treated subjects from the final matched sample.[12] Although NNM avoids bias due to incomplete matching, it can result in estimates that are more biased due to residual confounding than estimates produced using NNM with a caliper restriction.

In order to address the existence of bias due to residual confounding when using matching, several authors have described methods to combine matching-based methods with regression adjustment to reduce the effects of confounding. Rubin[13,14] considered using regression adjustment in matched samples when estimating treatment effects for continuous outcomes (ie, differences in means) and found that it reduced bias to a greater extent than either approach on its own. Rubin and Thomas[15] combined matching on the propensity score with adjustment for a limited set of prognostically important covariates when estimating treatment effects. These studies restricted their focus to settings with continuous outcomes. Although these studies examined the fitting of a linear regression model within the matched sample, other studies examined the use of regression models to impute potential outcomes, and concluded that the combination was superior to either method alone.[13,14,16,17] Belson[18] described a method for combining regression adjustment with matching that entailed fitting a linear regression model to the control subjects and using the fitted model effectively to estimate the potential outcome under control for the matched treated subjects. Gutman and Rubin[19] proposed a method for estimating the treatment log-odds ratio that combined multiple imputation (MI) with the use of two regression splines to impute the missing potential outcomes. Quade[20] described methods to combine matching with regression adjustment, while Imbens[16] described three different approaches to combine regression adjustment with matching on the propensity score when estimating treatment effects. Abadie and Imbens[21] developed the bias-corrected matching estimator that uses a regression model to impute the missing potential outcome using a regression-based approach. A recently proposed method called double-propensity score adjustment has been described that combines matching on the propensity score with covariate adjustment using the propensity score to impute potential outcomes under control in settings in which outcomes are continuous or binary.[22] More recently, Austin et al[23] examined the performance of a method to combine matching on the propensity score with regression adjustment when outcomes were times-to-event.

There are two primary limitations to many of the previous studies on combining matching with regression adjustment. First, many of the articles focus on continuous outcomes, whereas in biomedical and epidemiological research, binary outcomes (eg, disease remission or success vs failure of treatment) are also common.[24] Second, some of the suggested approaches involve using a regression model to impute a single potential outcome under control for each of the matched treated subjects. Often this involves a single application of conditional mean imputation (ie, the same value of the potential outcome under control would be imputed for all subjects with the same value of the covariates). A limitation of this approach is that the use of single imputation does not account for the uncertainty inherent in the imputed quantity, unless further adjustments are made. MI entails imputing multiple values for each missing value.[25-29] The use of MI allows the analyst to account for uncertainty in the imputed quantities when estimating estimands, such as risk differences. Furthermore, the use of MI may result in more efficient estimates. Although previous work has considered binary outcomes and imputed multiple potential outcomes under control, this study used the marginal odds ratio as the causal parameter.[19] From a clinical perspective, the risk difference (or its reciprocal, the number needed to treat) better informs medical decision making.[30,31] Conducting causal inference by simulating the missing potential outcomes from a Bayesian perspective was proposed by Rubin .[32,33]

The objective of the current study is to propose a new method for combining propensity-score matching with MI of potential outcomes under control for the matched treated subjects when estimating risk differences. The use of MI of potential outcomes under control is intended to serve the same function as the use of regression adjustment in matched samples: reducing the residual confounding that persists after the application of NNM. However, we anticipate that an advantage of proposed method is that it will result in improved sampling variance estimation due to explicit accounting of the uncertainty in the estimated potential outcome under control.

In Section 2, we describe a method for combining propensity-score matching with regression adjustment when outcomes are binary and describe its extension to imputing multiple potential outcomes under control for each matched treated subject. We then assess the performance of our proposed method. In particular, it is important to assess: (i) the effect of the number of multiply imputed potential outcomes under control on the resulting quality of inferences; (ii) the performance of the proposed method compared with that of conventional NNM; (iii) the performance of the proposed method with that of conventional NNM combined with *single* imputation of the potential outcomes under control. Accordingly, in Section 3, we describe simulations to conduct these assessments, and in Section 4 we report the results of these simulations. In Section 5, we provide a brief case study illustrating the application of the proposed method. In Section 6, we summarize our findings and place them in the context of the existing literature.

2 | METHODS FOR COVARIATE-ADJUSTED ANALYSES IN PROPENSITY-SCORE MATCHED SAMPLES WITH BINARY OUTCOMES

2.1 | Notation

The observed covariates for a subject are denoted by $\mathbf{X}$. The sample has N_t treated and N_c control subjects. The matching method evaluated here has one matched control for each treated subject, so that $N_{mc} = N_t$, but the method can be applied with minor modifications when there is more than one match per treated subject. Each subject has a binary variable, Z, indicating whether the subject received the control ($Z = 0$) or treatment ($Z = 1$) intervention. It is assumed that each subject has a probability of receiving the treatment intervention conditional on their measured covariates. These are the propensity scores denoted by $\Pr(Z = 1|\mathbf{X})$. It is also assumed that the sample, $\mathbf{X}$, was drawn from a distribution denoted by $\mathbf{F}$. This implies conditional distributions for $\mathbf{X}$ given $Z = 0$, and $Z = 1$, which are denoted by $\mathbf{F}_C$ and $\mathbf{F}_T$.

Each subject has two potential binary outcomes, denoted by Y_t and Y_c, which would be observed when the subject receives the treatment or control intervention, respectively. Only one of the two potential outcomes is observable. The potential outcomes have probability distributions given baseline covariates, denoted by $\mathrm{P}(Y_c = 1|\mathbf{X})$ and $\mathrm{P}(Y_t = 1|\mathbf{X})$. The probability of a positive outcome to the treatment among subjects assigned to the treatment is given by $\mathrm{P}_t = \int P(Y_t = 1|\mathbf{X})dF_t(\mathbf{X})$. The probability of a positive outcome from the control intervention, if it could be observed among subjects assigned to the treatment intervention, is denoted by $\mathrm{P}_t^c = \int P(Y_c = 1|\mathbf{X})dF_t(\mathbf{X})$. Methods in the remainder of Section 2 provide estimators for P_t and P_t^c, with an emphasis on the risk difference, $P_t - P_t^c$.

2.2 | Nearest neighbor matching on the estimated propensity score

The propensity scores will be estimated by expanding the model to include an unknown parameter assuming the treatment assignment variables are independent conditional on the covariates and the unknown parameter. A simple linear logistic model is used:

$$\mathrm{P}(\mathrm{Z} = 1|\mathbf{X}, \beta_{0,\mathrm{treat}}, \boldsymbol{\beta}_{\mathrm{treat}}) = \mathrm{logit}^{-1}(\beta_{0,\mathrm{treat}} + \mathbf{X}\boldsymbol{\beta}_{\mathrm{treat}}). \quad (1)$$

Maximum likelihood (ML) estimation is used to estimate $\beta_{0,\mathrm{treat}}, \boldsymbol{\beta}_{\mathrm{treat}}$ by $\widehat{\beta}_{0,\mathrm{treat}}, \widehat{\boldsymbol{\beta}}_{\mathrm{treat}}$. The propensity scores are estimated using the $\widehat{\beta}_{0,\mathrm{treat}}, \widehat{\boldsymbol{\beta}}_{\mathrm{treat}}$ in (1) applied to both the treated and control subjects.

The NNM matching was implemented by randomly ordering treated subjects and then matching the control subject whose propensity score was closest to that of each treated subject. This process was repeated until a matched control subject had been identified for each treated subject (we assume that the reservoir of control subjects is larger than the number of treated subjects). Matching was done without replacement, so each control subject could be selected at most once. The outcome proportions from the treated and matched control samples are denoted by $\overline{Y}_t, \overline{Y}_{mc}$, and $\overline{Y}_t - \overline{Y}_{mc}$ denotes the risk difference, which are estimators of P_t and P_t^c, and P_t-P_t^c, respectively. The ignorability condition required for $\overline{Y}_{mc}$ to estimate consistently P_t^c is given in Rosenbaum and Rubin,[1] so it is not repeated here.

The estimator of a standard error (SE) for $\overline{Y}_t - \overline{Y}_{mc}$, recommended for use with propensity-score matched samples[34], was derived for use with McNemar's test. Let b denote the number of matched pairs in which the treated subject experienced the outcome and the matched control subject did not, whereas c denotes the number of matched pairs in which the treated subject did not experience the outcome and the matched control subject did. Then the estimated SE of the

estimated risk difference is $\sqrt{\frac{(b+c)-(c-b)^2/N_t}{N_t^2}}$.[35] Note that McNemar's formula is a specialization of the usual SE for paired comparisons.

2.3 | A covariate-adjusted matched estimator using MI

The NNM matching on the estimated propensity scores in Section 2.2 reduces the bias in the naïve estimator $\overline{Y}_t - \overline{Y}_c$. Even when the ignorability condition assures that there is no bias due to unmeasured covariates, the matched sample mean may still be conditionally biased when the matching is not exact, or the estimated propensity score model does not accurately represent the relationship between treatment assignment and the observed covariates. A simple method to adjust $\overline{Y}_{mc}$ for any remaining difference between the covariate distributions in the treated and matched control subjects is presented here.

A logistic regression model is assumed relating the outcome to the baseline covariates when subjects receive the control intervention:

$$\mathrm{P}(Y_c = 1|\mathbf{X}, \beta_{0,\text{outcome}}, \boldsymbol{\beta}_{\text{outcome}}) = \text{logit}^{-1}(\beta_{0,\text{outcome}} + \mathbf{X}\boldsymbol{\beta}_{\text{outcome}}). \tag{2}$$

As with the propensity score estimation, ML estimation yields estimators $\hat{\beta}_{0,\text{outcome}}, \hat{\boldsymbol{\beta}}_{\text{outcome}}$, and their standard asymptotic variance-covariance matrix, denoted by $\hat{\Sigma}$. A key feature of this method is that the estimation is applied to the matched control sample, rather than the full control sample. The estimated conditional response probabilities for the matched control subjects are denoted by $\hat{P}_{1mc}, \ldots, \hat{P}_{n_t mc}$. Because the logistic regression model includes an intercept, the mean of the estimated response probabilities is

$$\frac{1}{N_t}\sum_{i=1}^{N_t}\hat{P}_{imc} = \overline{Y}_{mc}. \tag{3}$$

The estimator in (3) can now be adjusted for any remaining differences in covariates by computing and averaging the estimated conditional response probabilities for the treated subjects rather than the matched control subjects. The estimated response probabilities for each treated subject, had they received the control intervention, are denoted by $\hat{P}^c_{1t}, \ldots, \hat{P}^c_{N_{tt}}$ (with $N_{t\ t}$ denoting the number of treated subjects), and their mean is given by

$$\bar{\hat{P}}^c_t = \frac{1}{N_t}\sum_{i=1}^{N_t}\hat{P}^c_{it}. \tag{4}$$

The estimator in (4) coincides with $\overline{Y}_{mc}$ when the treated and matched control covariates have the same distribution. Provided the propensity score matching is successful, $\bar{\hat{P}}^c_t$ makes small adjustments to $\overline{Y}_{mc}$ by making small changes to the covariate distribution evaluated. Note that a similar derivation can be applied if the estimation is applied to the full control sample and the estimated response probabilities are computed for all control subjects. In this case the adjustment is to the naïve estimator $\overline{Y}_c$, and the adjustment may be large if the original (ie, prematching) treated and control covariate distributions differ substantially. This introduces more reliance on correct specification of the response model, which is further exacerbated because the model estimates may be dominated by a few control subjects that are very different from any treated subject.

MI provides a simple approach to estimate the SE for $\bar{\hat{P}}^c_t$ and the risk difference $\overline{Y}_t - \bar{\hat{P}}^c_t$. To create M imputed datasets, draw new estimated regression parameters from their asymptotic distribution

$$(\beta_{0,\text{outcome}}, \boldsymbol{\beta}_{\text{outcome}}) \sim N((\hat{\beta}_{0,\text{outcome}}, \hat{\boldsymbol{\beta}}_{\text{outcome}}), \hat{\Sigma}). \tag{5}$$

A set of control response probabilities are computed for the treated subjects using each of the M simulated $(\beta_{0,\text{outcome}}, \boldsymbol{\beta}_{\text{outcome}})$ parameters, and these response probabilities are averaged over the treated subjects yielding M estimators of the form $\overline{P}^c_t$ and $\overline{Y}_t - \overline{P}^c_t$. The usual approach to inference with multiply imputed estimators can then be

TABLE 1 Factors in the Monte Carlo simulation

Factor	Levels
Risk differences	0, −0.01, −0.02, −0.03, −0.04, −0.05
Prevalence of treatment	0.05, 0.10, 0.15, 0.20, 0.25
Strength of treatment-selection and outcome models	Weak, moderate, strong

applied. The complete-data variance of the risk difference estimator is the square of the usual SE for a paired comparison (eg, pairs Y_{it} and P^c_{it}).

It natural to consider the potential 0/1 control outcomes, Y^c_{it}, as missing data so they are imputed rather than P^c_{it}. The Y^c_{it} can be easily generated from the P^c_{it}. For large M, this will yield asymptotically equivalent MI estimators, as in the previous paragraph. The SE computed using the usual MI methods applied to imputed control outcomes are too large, however, because the simple complete-data sample mean estimator may not be congenial with the logistic regression imputation model.[36] The complete-data analysis based on the imputation model would fit the logistic regression to both the matched control and the control outcomes of the treated subjects, which can be more efficient than using only the control responses of the treated subjects.

Although the regression model is used to generate the model for imputing potential outcomes under control, we will use the terminology of "combining regression adjustment with propensity score matching" to reflect the intent of our method, which is to remove some of the effects of residual confounding that persist after matching alone. Thus, we will speak of an adjusted risk difference.

3 | MONTE CARLO SIMULATIONS TO EXAMINE THE EFFECT OF THE NUMBER OF IMPUTED OUTCOMES UNDER CONTROL ON ESTIMATION AND INFERENCE ABOUT THE RISK DIFFERENCE

Monte Carlo simulations were used to examine the effect of the number of imputed potential outcomes under control for each matched treated subject on estimation and inference about the treatment risk difference. We conducted a set of simulations in which three population factors were studied: (i) the true risk difference relating treatment (Z) to the probability of the outcome (Y); (ii) the prevalence of treatment; (iii) the strength of the treatment-selection process. Other factors were fixed in the design of the simulation: the number of baseline covariates, their distribution, and the distribution of outcomes. We also compared the performance of the proposed method with different numbers of MIs (including the special case of single imputation) and with that of conventional NNM. The factors of the Monte Carlo simulations and the different levels of each factor are summarized in Table 1.

3.1 | Empirical analyses to inform the design of the Monte Carlo simulations

We conducted a series of empirical analyses to obtain parameter values that were used to generate data for the simulations. These empirical analyses were conducted using data from a previously published study comparing the balancing properties of different propensity score methods.[37] The sample consisted of 7101 patients discharged from hospital with a diagnosis of heart failure in Ontario, Canada. These data were collected as part of the Enhanced Feedback for Effective Cardiac Treatment Study, a study designed to improve the quality of care provided to patients with cardiovascular disease.[38] The exposure of interest Z in the current set of analyses was whether the patient received such a prescription. Of the 7101 patients, 1895 (26.7%) received a beta-blocker prescription at hospital discharge. The outcome ($Y = 1$) was death within 1 year of hospital discharge. No patients were censored prior to 1 year of discharge. Two thousand and seventy-six (29.2%) patients died within 1 year of discharge (ie, had $Y = 1$).

For the purposes of these simulations, we used 10 baseline covariates: age, systolic blood pressure at admission, heart rate at admission, respiratory rate at admission, initial laboratory value of sodium, history of previous myocardial infarction, history of peripheral arterial disease, history of chronic obstructive pulmonary disease, history of dementia, and history of cancer. The first five are continuous, whereas the last five are binary. We standardized the five continuous variables so that they had mean zero and unit variance. We denote the sample variance-covariance matrix of these 10

variables by $\boldsymbol{\Sigma}_{\text{empirical}}$, which was used to generate data for the simulations, so that the simulated baseline variables will have a correlation structure that resembles the empirical data.

These 10 covariates were selected because logistic regression models identified them as being associated with both the log-odds of beta-blocker use (the treatment) and with the odds of death within 1 year (the outcome Y). Let $\widehat{\boldsymbol{\gamma}}_{\text{treat}}$ and $\widehat{\boldsymbol{\gamma}}_{\text{outcome}}$ denote the ML estimated vectors of regression coefficients for the 10 covariates from the logistic regression models for treatment and mortality, respectively (ie, the intercept term has been removed). These two vectors will be used as the population parameters when simulating treatment status and the outcome so that the association between baseline covariates and both the exposure and the odds of the outcome in the simulated data resemble the empirical data. Note that we use γ's to denote the regression coefficients from these empirical regression equations to distinguish them from those described in the previous section that are used for estimating the propensity score model and the outcomes model applied to the resulting simulated data.

3.2 | Simulating baseline covariates and treatment status

We simulated baseline covariates, treatment status, and a binary outcome for a superpopulation consisting of 1 000 000 subjects. Ten baseline covariates $(X_1, \ldots, X_{10})$ were simulated for each subject from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance $\boldsymbol{\Sigma}_{\text{empirical}}$. The first five covariates $(X_1, \ldots, X_5)$ were retained as continuous covariates. The last five covariates were dichotomized using an appropriate percentile so that the prevalence of each of the five binary covariates was the same as that of one of the binary covariates used in the empirical analyses described above. Thus, the simulated covariates had a multivariate structure similar to that observed in the data described above.

The logit of the probability of treatment was determined for the ith subject using the following logistic model: $\text{logit}(p_i) = \gamma_{0,\text{treat}} + \widehat{\boldsymbol{\gamma}}_{\text{treat}}\mathbf{X}_i$, where $\mathbf{X}_i$ is the vector of 10 baseline covariates for the ith subject. For each subject, treatment assignment was simulated from a Bernoulli distribution with subject-specific parameter p_i. The intercept of the treatment-selection model $(\gamma_{0,\text{treat}})$ was determined using a bisection approach so that the prevalence of treatment is equal to a specified value (see Table 1 for prevalences of treatment).

3.3 | Simulating a binary outcome

For each subject in the large superpopulation of 1 000 000 subjects, we simulated two binary potential outcomes: a binary outcome under control and a binary outcome under treatment. The linear predictor under control was defined as: $\text{LP}_i(0) = \gamma_{0,\text{outcome}} + \mathbf{X}_i\widehat{\boldsymbol{\gamma}}_{\text{outcome}}$. We used a bisection approach to determine $\gamma_{0,\text{outcome}}$ so that the proportion of treated subjects experiencing a positive outcome under control was 0.20 (this was fixed by design).

The linear predictor under treatment is $\text{LP}_i(1) = \text{LP}_i(0) + \log(\gamma_{\text{treat}})$. We used a bisection approach to determine the value of γ_{treat} so that the marginal risk difference in the treated subjects in the superpopulation was the desired quantity (see below). The marginal risk difference among the treated subjects was determined as the mean difference in potential outcomes in the treated subjects in the superpopulation.

3.4 | Monte Carlo simulations

From the superpopulation, we drew a random sample of 10 000 without replacement. In this random sample, we estimated the propensity score using logistic regression and created a matched sample using NNM to match treated and control subjects on the estimated propensity scores. We then conducted the simple matched comparison analysis described in Section 2.2, and the analysis of the covariate-adjusted matched estimator in Section 2.3. The multiply imputed covariate-adjusted estimator was computed for differing numbers of imputations ranging from $M = 1$ to $M = 100$ in increments of 1.

From the superpopulation, we drew 1000 random samples without replacement, each of size 10 000, and conducted the statistical analyses described above. The impact of the number of imputed adjusted potential outcomes under control for each matched subject on the precision of the estimated treatment effect was assessed in the following ways: (i) the SD of the estimated risk difference across the 1000 simulation replications was computed for each value of M ($M = 1, 2, \ldots, 100$). This analysis examines the extent to which increasing the number of imputed potential outcomes under

control reduces the sampling variability of the estimated risk difference; (ii) the mean estimated SE of the estimated risk difference was computed across the 1000 simulation replications for each scenario and for each value of M. The ratio of the mean estimated SE with $M = 1$ vs that with $M > 1$ for each $M > 1$ was computed; the mean estimated SEs were also compared with the SD of the empirical sampling distribution of the risk differences across the 1000 simulation replicates (iii) for each scenario and for each value of M, we computed the proportion of estimated 95% confidence intervals for the estimated risk difference that contained the true risk difference; (iv) for each scenario and for each value of M, we computed the mean squared error (MSE) of the estimated risk difference.

Three population factors varied in the Monte Carlo simulations (refer to Table 1): (i) the true risk difference for treatment; (ii) the prevalence of treatment; (iii) the strengths of the covariate predictors in the treatment-selection process and in the outcomes model. The true risk difference took six values as given in Table 1. The prevalence of treatment took five values as given in Table 1. We considered treatment-selection and outcome processes of three different strengths. First, we used the empirically estimated vectors $\boldsymbol{\gamma}_{\text{treat}}$ and $\boldsymbol{\gamma}_{\text{outcome}}$ in the data-generating process when simulating treatment status and outcomes, respectively. We then repeated this process using $2\boldsymbol{\gamma}_{\text{treat}}$ and $2\boldsymbol{\gamma}_{\text{outcome}}$ and $3\boldsymbol{\gamma}_{\text{treat}}$ and $3\boldsymbol{\gamma}_{\text{outcome}}$. We refer to the three levels of this factor as weak, moderate, and strong treatment-selection and outcome processes, respectively. We thus constructed 90 (6 risk differences $\times$ 5 prevalence of treatment $\times$ 3 strengths of treatment-selection and outcome processes) superpopulations of subjects.

3.5 | Description of simulated datasets

We briefly describe the degree of imbalance in baseline covariates between treated and control subjects in the superpopulation across the different scenarios of the simulations. The relationship between the prevalence of treatment, the strength of the treatment-selection model and the Mahalanobis distance between the treated subjects and the control subjects in the entire superpopulation is described in Figure 1 (solid black line with solid black circles plotting symbols). The Mahalanobis distance between treatment and control groups increases with the strength of the treatment-selection process. When the strength of the treatment-section process was weak, the Mahalanobis distance was approximately equal to 0.25. When the strength of the treatment-section process was moderate, the Mahalanobis distance ranged from 0.74 to 0.83. When the strength of the treatment-section process was strong, the Mahalanobis distance ranged from 1.24 to 1.55.

The relationship between the prevalence of treatment, the strength of the treatment-selection process and the absolute standardized mean difference comparing the mean of each of the 10 baseline covariates between treated and control subjects in the superpopulation is described also in Figure 1.[10,39] There is one line for each of the 10 covariates. As the strength of the treatment-selection process increased, the number of baseline covariates with a large standardized mean difference increased. The Mahalanobis distance decreased with increasing prevalence of treatment (with moderate or strong treatment-selection processes).

When the prevalence of treatment was 0.05, the mean number of treated subjects in each simulated dataset was approximately 500, and the mean number of potential controls for each treated subject was approximately 19. When the prevalence of treatment was 0.25, the mean number of treated subjects in each simulated dataset was approximately 2500, and the mean number of potential controls for each treated subject was approximately 3.

3.6 | Description of fraction of missing information across the simulation scenarios

The fraction of missing information (FMI) is computed as

$$\text{FMI} = 1 - \frac{\text{Variance of complete-data estimator}}{\text{Variance of the MI estimator}}.$$

The "complete data" consist of both outcomes for the treated subjects, and the conditional probabilities for the outcomes of the treated subjects to the control intervention. The variance of the complete-data estimator is the average across the M imputations of the variance for the pairwise mean difference of the Y_{it} and P_{it}^{c} pairs. We then computed the mean of the complete-data variances across the simulation replicates. The MI-based sampling variances of the estimator were also averaged across the simulation replicates. For the purposes of estimating FMI, we used 10 000 simulation replicates

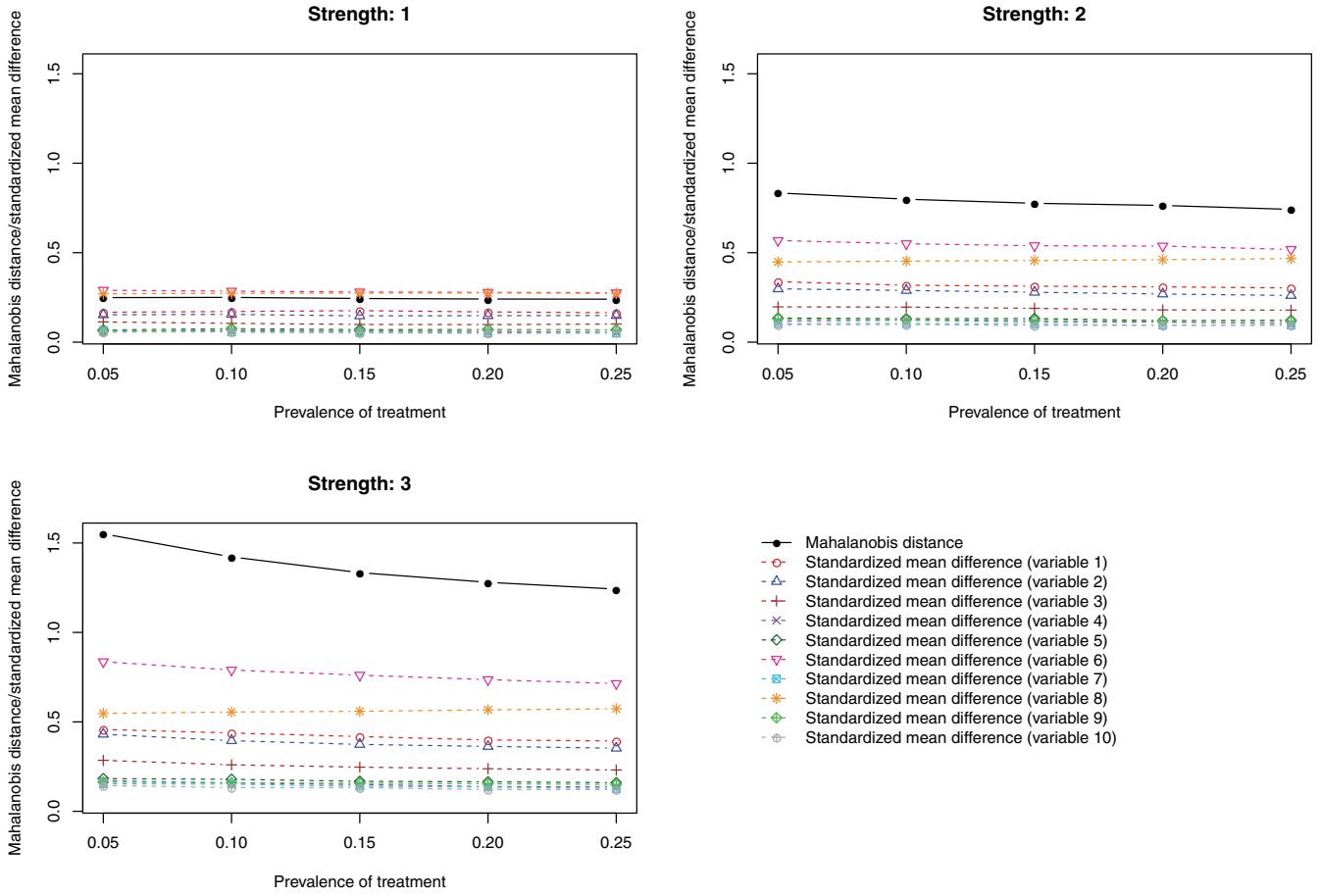

FIGURE 1 Balance of baseline covariates between treatment groups [Colour figure can be viewed at wileyonlinelibrary.com]

and $M = 100$ imputed potential outcomes under control. Across the 90 scenarios the FMI ranged from 0.49 to 0.55, with a median of 0.52 (25th and 75th percentiles: 0.51 and 0.54), which is close to its anticipated value of 0.50.

4 | RESULTS OF THE MONTE CARLO SIMULATIONS

4.1 | Effect of number of imputations on the sampling variability of estimated risk differences

The relationship between M (the number of imputed potential outcomes under control) and the SD of the empirical sampling distribution of the estimated risk differences are presented in Figure 2 for five scenarios (true risk difference = −0.02 and the moderate treatment-selection process, with prevalence of treatment ranging from 0.05 to 0.25 in increments of 0.05). There is one curve for each of the five prevalences of treatment. On the figure we have superimposed two vertical lines denoting $M = 5$ and $M = 20$. For each prevalence of treatment, there is an initial rapid decrease in the sampling variability of the estimated risk difference as M increases from 1 to 5. Once M is greater than approximately 20, the curves become approximately flat, and any subsequent decrease in sampling variability is negligible. On each curve we have superimposed a fitted linear model of the form $a + b/\sqrt{M}$. The R^2 statistics of the five linear models ranged from 0.90 to 0.92.

The relative change in the SD of the sampling distribution when using $M = 5$ and $M = 20$ compared with when a single potential outcome under control is imputed is described in Figure 3 for all 90 scenarios (ie, we divided the sampling variability of the risk difference when $M = 5$ or $M = 20$ by the sampling variability of the risk difference when $M = 1$). The figure consists of two panels, one for each of the two values of M. Each panel consists of a series of dot charts, with

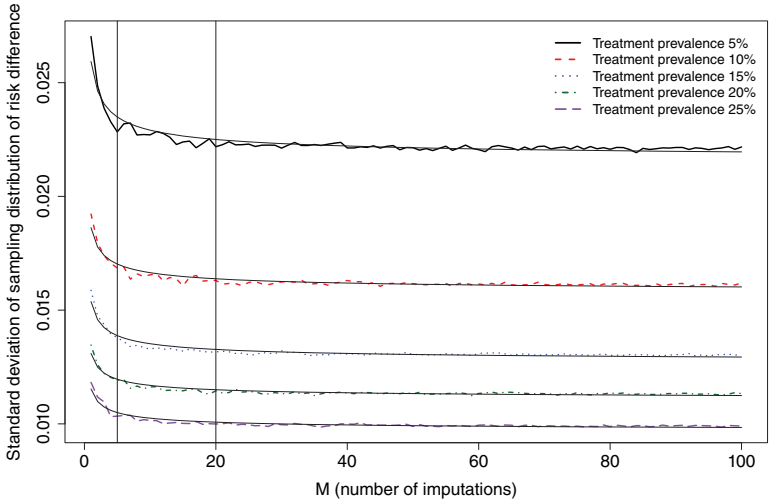

FIGURE 2 Empirical SD of sampling distribution of risk difference as a function of number of imputations [Colour figure can be viewed at wileyonlinelibrary.com]

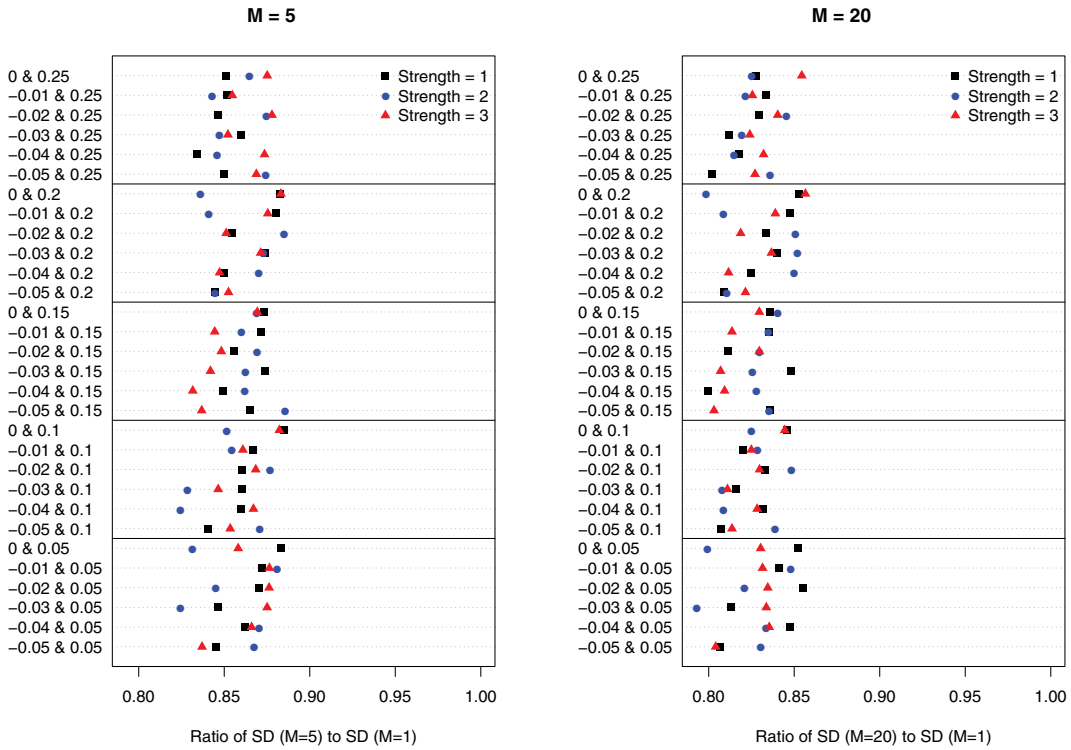

FIGURE 3 Relative change in SD of sampling distribution of estimated risk difference with $M = 5$ or 20 compared with $M = 1$ [Colour figure can be viewed at wileyonlinelibrary.com]

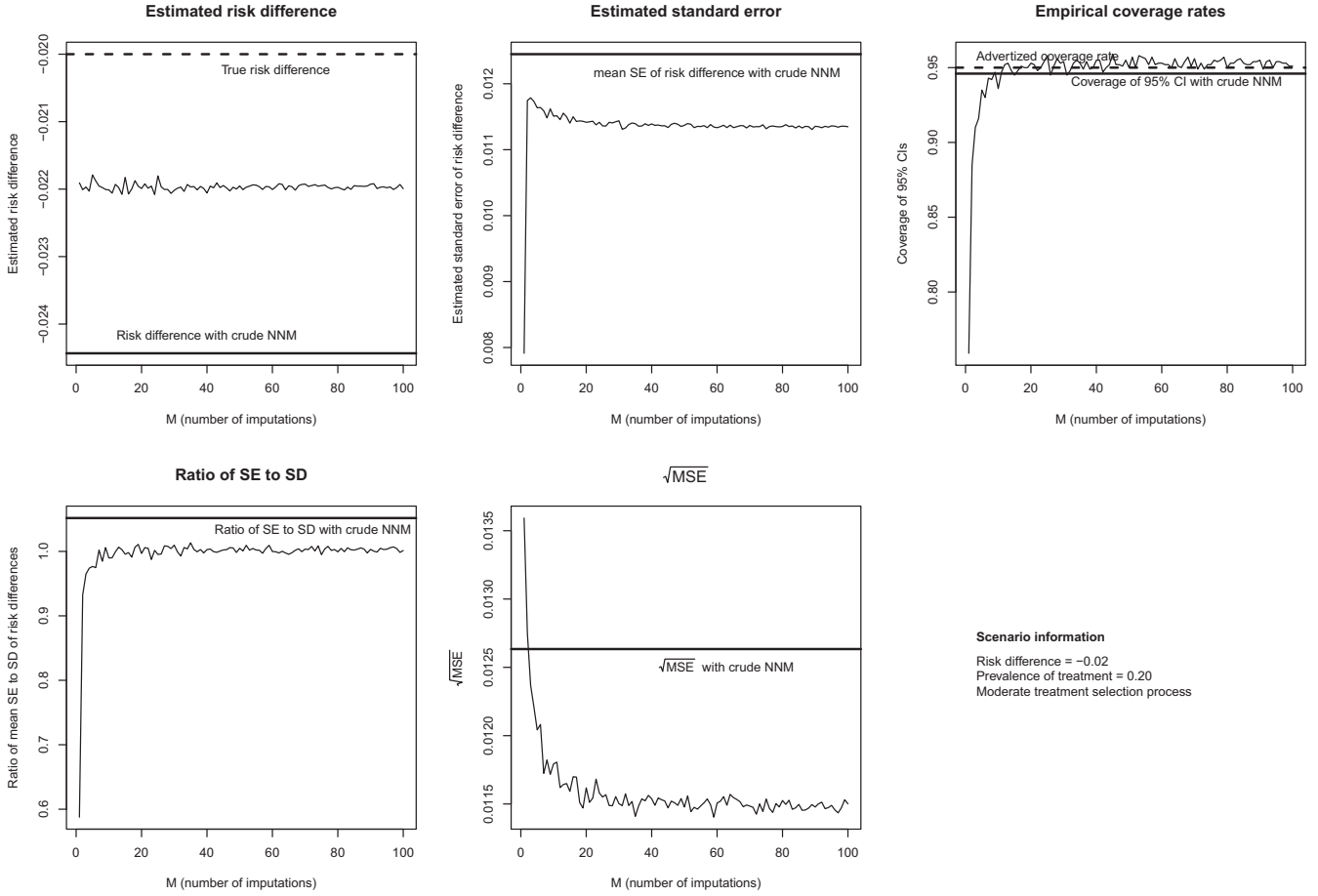

FIGURE 4 Effect of number of imputations (M) on inferences in one scenario

one horizontal line for each combination of true conditional risk difference and prevalence of treatment. The first set of three scenarios (reported on the top horizontal line of the dot chart), for example, is labeled "0 & 0.25" to indicate that in these scenarios the true risk difference was 0 and the prevalence of treatment was 0.25. The other rows are labeled in a similar fashion. On each horizontal line are three dots, one for each of the three strengths of the treatment-selection model. The location of each dot denotes the relative reduction in the sampling variation in the risk difference when the given number of imputed outcomes under control were used ($M = 5$ and 20). Using $M = 5$ reduced the SD of the sampling distribution by between 11% and 18% compared with when using $M = 1$. Using $M = 20$ reduced the SD of the sampling distribution by between 14% and 21% compared with when using $M = 1$.

4.2 | Initial exploration of relationship between M and inferences about the risk difference

We restricted our initial examination to one scenario (true risk difference = −0.02, prevalence of treatment = 0.20, and moderates strength treatment-selection process). The relationship between M and inferences about the risk difference are presented in Figure 4.

The relationship between M and the estimated risk difference is described in the top left panel. On this panel we have superimposed two horizontal lines. The upper horizontal line denotes the true risk difference (dashed horizontal line) and the lower horizontal line denotes the mean estimated risk difference obtained using crude NNM (solid horizontal line). Although the true risk difference was −0.02, the mean estimated risk difference using crude NNM was approximately −0.024, for a relative bias of approximately 20%. By contrast, the estimated risk difference when imputing adjusted potential outcomes under control displayed decreased bias, with a relative bias of approximately 10%. As would be expected,

the number of imputed outcomes under control does not affect the magnitude of bias (the small variation in the estimated risk difference across the different values of M simply reflects simulation error).

The relationship between M and the estimated SE of the risk difference is described in the top center panel. On this panel we have superimposed a horizontal line denoting the mean SE of the estimated risk difference when using crude NNM. When using single imputation ($M = 1$), the mean estimated SE was substantially lower than that obtained using crude NNM. However, once $M \geq 2$, the mean estimated SE of the risk difference was larger than obtained when using single imputation ($M = 1$), but smaller than that obtained using crude NNM. When using $M \geq 2$, there was an initial decrease in the mean estimated SE as M increased. However, once M exceeded 20, the marginal decrease in the SE was negligible, consistent with theoretical results from MI in general.[25]

The relationship between M and the empirical coverage rates of 95% confidence intervals is described in the top right panel. On this panel we have superimposed two horizontal lines. The upper dashed horizontal line denotes the advertised coverage rate of 0.95, whereas the lower solid horizontal line denotes the empirical coverage rate when using crude NNM. The empirical coverage rate when using crude NNM was close to the nominal 0.95. When using single imputation ($M = 1$), the empirical coverage rate was substantially lower than nominal, as expected. When using MI of adjusted outcomes under control, with $M \geq 10$, the empirical coverage rate was approximately equal to the nominal rate.

The relationship between M and the ratio of the mean estimated SE to the SD of the empirical sampling distribution of the risk difference is described in the bottom left panel. On this panel we have superimposed a horizontal line denoting this ratio when using crude NNM. We observe that the ratio is closer to unity when imputing potential outcomes under control than when using crude NNM once M is approximately greater than 5. When using single imputation ($M = 1$), the ratio is approximately equal to 0.60, indicating that the estimated SE underestimates the SD of the sampling distribution by about 40%. Once $M \geq 4$, the ratio ranges from 0.97 to 1.01.

The relationship between M and the square root of MSE of the estimated risk difference is described in the bottom center panel (we report the square root of MSE to be consistent with our reporting of SEs and SDs elsewhere). On this panel we have superimposed a horizontal line showing the square root of MSE when using crude NNM. When using single imputation or MI with $M = 2$, the square root of MSE of the estimated risk difference was greater than the square root of the MSE of the risk difference when using crude NNM. With $M > 2$, the square root of MSE was smaller than that obtained when using crude NNM. Once M was greater than approximately 40, increasing M had a negligible effect on the square root of MSE.

Based on these initial analyses, in the subsequent sections we examine inferences about the risk difference when imputing the adjusted potential outcome under control using $M = 1$ and $M = 20$. We compare these inferences with those made using conventional crude NNM methods.

4.3 | Bias and relative bias of the estimated risk difference

The bias in the estimated risk difference across the 90 scenarios is reported in Figure 5. There is one panel for each of the three strengths of the treatment-selection process. On each panel we report the bias when using crude NNM (solid lines), adjusted potential outcomes with $M = 1$ (dashed lines), and adjusted potential outcomes with $M = 20$ (dotted lines). Relative bias is similarly reported in Figure 6 (with the omission of the 15 scenarios in which the true risk difference was equal to zero, as the relative bias is not defined in these scenarios). Note that bias should be independent of M. The observed differences between different values of M are negligible (as expected) and due to simulation error. As the strength of the treatment-selection process increased, the prevalence of treatment had a more pronounced effect on bias when using crude NNM, with bias increasing with increasing prevalence of treatment when the strength of the treatment-selection process was moderate or strong. By contrast, when imputing potential outcomes under control (using either $M = 1$ or $M = 20$), bias was minimal, except when the prevalence of treatment was very low. As expected, there were no apparent differences in bias between using $M = 1$ and $M = 20$.

4.4 | Estimation of SEs of the estimated risk difference

The mean estimated SEs of the estimated risk differences across the 90 scenarios are reported in Figure 7, which has a similar structure to the previous two figures. On each panel we have superimposed a solid black curve denoting the

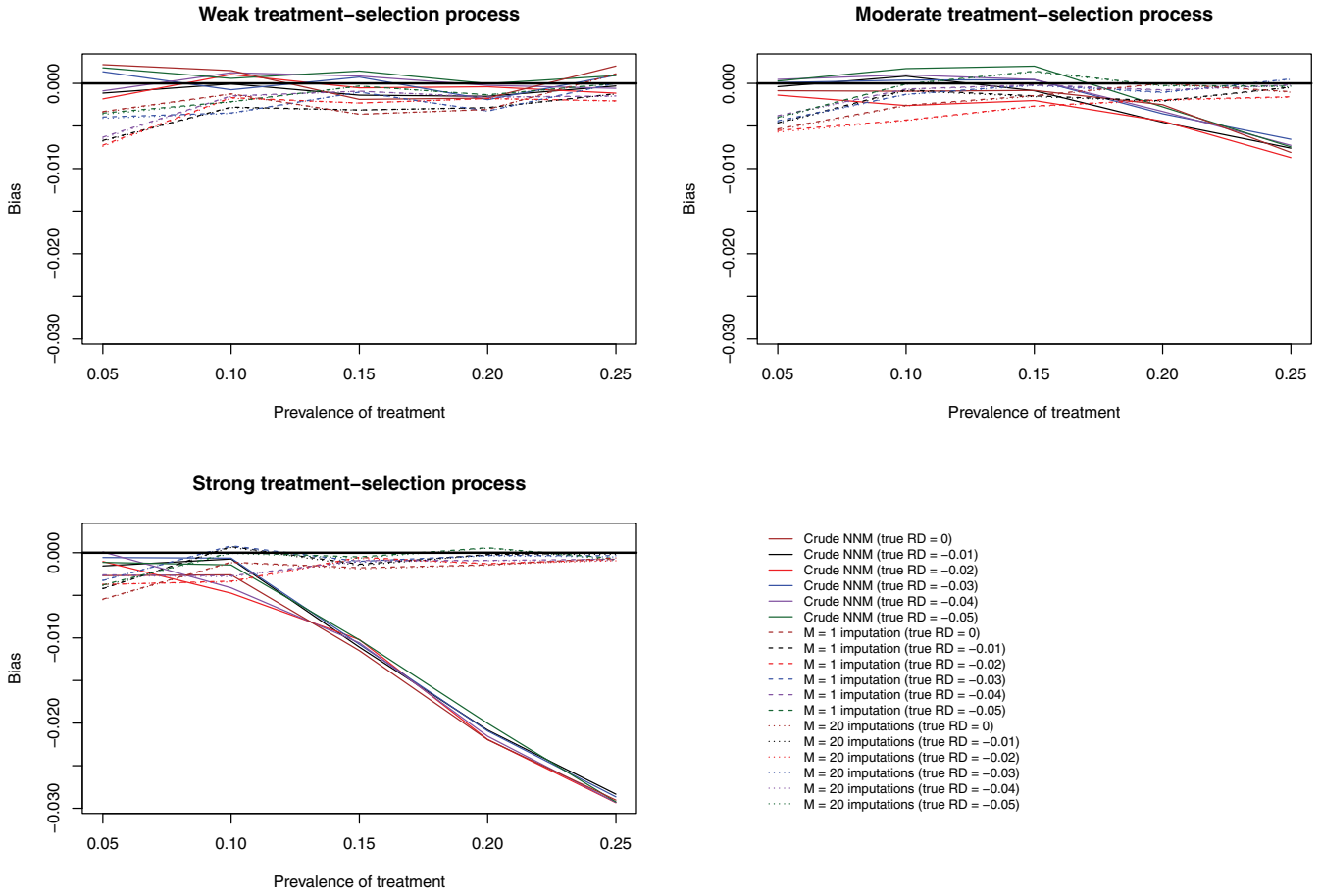

FIGURE 5 Bias in risk difference [Colour figure can be viewed at wileyonlinelibrary.com]

relationship $1/\sqrt{2p10000}$, where p denotes the prevalence of treatment (thus $2p10000$ denotes the number of subjects in the matched sample constructed using NNM).

Regardless of the estimation method (crude NNM, imputation with $M = 1$ or 20), the mean estimated SEs decreased with increasing prevalence of treatment, reflecting the increased sample size of the matched sample. The estimated SEs tended to be largest when using crude NNM and smallest when using single imputation of potential outcomes under control. As the strength of the treatment-selection process increased, differences between the use of crude NNM and MI with $M = 20$ increased, with the SEs of the latter becoming smaller than those of the former. By comparing the curves to the superimposed relationship described above, one notes that the decrease in SEs with increasing prevalence of treatment is consistent with the increasing size of the matched samples.

4.5 | MSE of estimated risk difference

The square root of the MSE of the estimated risk difference across the 90 scenarios are reported in Figure 8 (we report the square root of the MSE to be consistent with our reporting of SEs and SDs). MSE tended to decrease with increasing prevalence of treatment (reflecting the increased size of the matched sample), with an increase when using NNM when there was strong treatment-selection process and the prevalence of treatment was 0.15 or greater (reflecting the previously observed increase in bias in these scenarios). MSE was always higher when using single imputation of the potential outcomes under control than when $M = 20$ outcomes under control were imputed. MSE was comparable between crude NNM and MI with $M = 20$, except when there was a strong treatment-selection process, in which case the latter approach resulted in estimates with lower MSE.

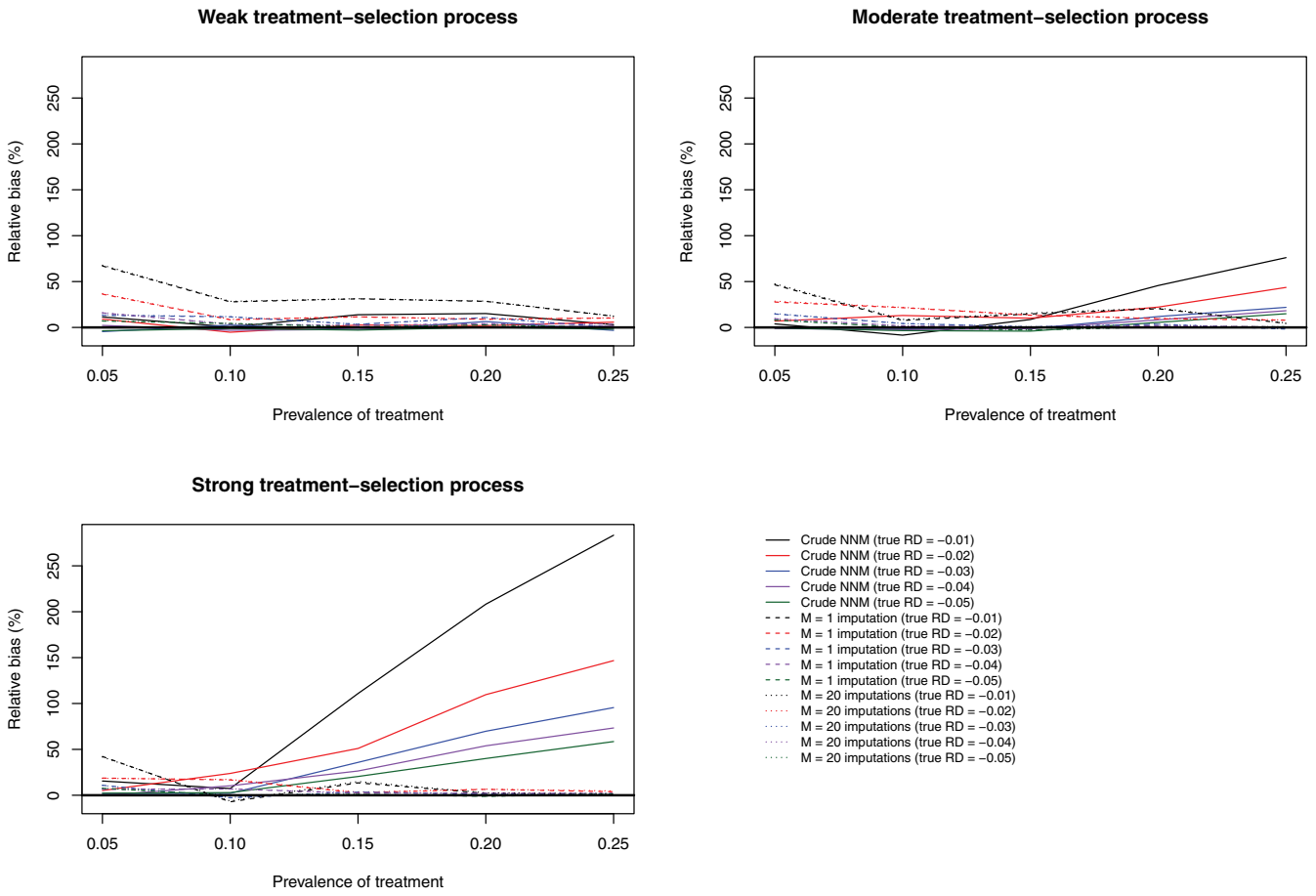

FIGURE 6 Relative bias in risk difference [Colour figure can be viewed at wileyonlinelibrary.com]

4.6 | Coverage of confidence intervals

The empirical coverage rates of estimated 95% confidence intervals across the 90 scenarios are reported in Figure 9. On each panel we have superimposed three horizontal lines denoting the nominal coverage rate (0.95) and rates of 0.936 and 0.964. Due to our use of 1000 simulation replicates, empirical coverage rates that lie between the latter two rates are not statistically significant from the advertised rate of 0.95 using standard normal-theory. The use of crude NNM estimation resulted in confidence intervals that had approximately the correct coverage rates when the strength of the treatment-selection process was weak or moderate. However, when the strength of this process was strong and the prevalence of treatment was at least 0.15, empirical coverage rates were sub-optimal, consistent with the bias in the estimated risk difference (see Figures 5 and 6). Single imputation of the potential outcome under control tended to result in 95% confidence intervals whose empirical coverage rates were less than 0.80. By contrast, MI of the potential outcomes under control (using $M = 20$) resulted in 95% confidence intervals whose coverage rates tended to equal 95%.

4.7 | Ratio of mean of estimated SEs to the SD of the sampling distribution

The ratios of the mean estimated SE to the SD of the empirical sampling distribution of the estimated risk difference across the 90 scenarios are reported in Figure 10. On each panel we have superimposed a horizontal line denoting a ratio of one. Points above this line denote that the estimated SE overestimates the SD of the sampling distribution of the risk difference. Points below this line indicate that the estimated SE underestimates the SD of the sampling distribution of the risk difference. When the strength of the treatment-selection process was weak, the use of crude NNM estimation resulted in SEs that correctly approximated the SD of the sampling distribution of the risk difference. As the strength of the treatment-selection process increased, the ratio for NNM increased. However, even when the strength of

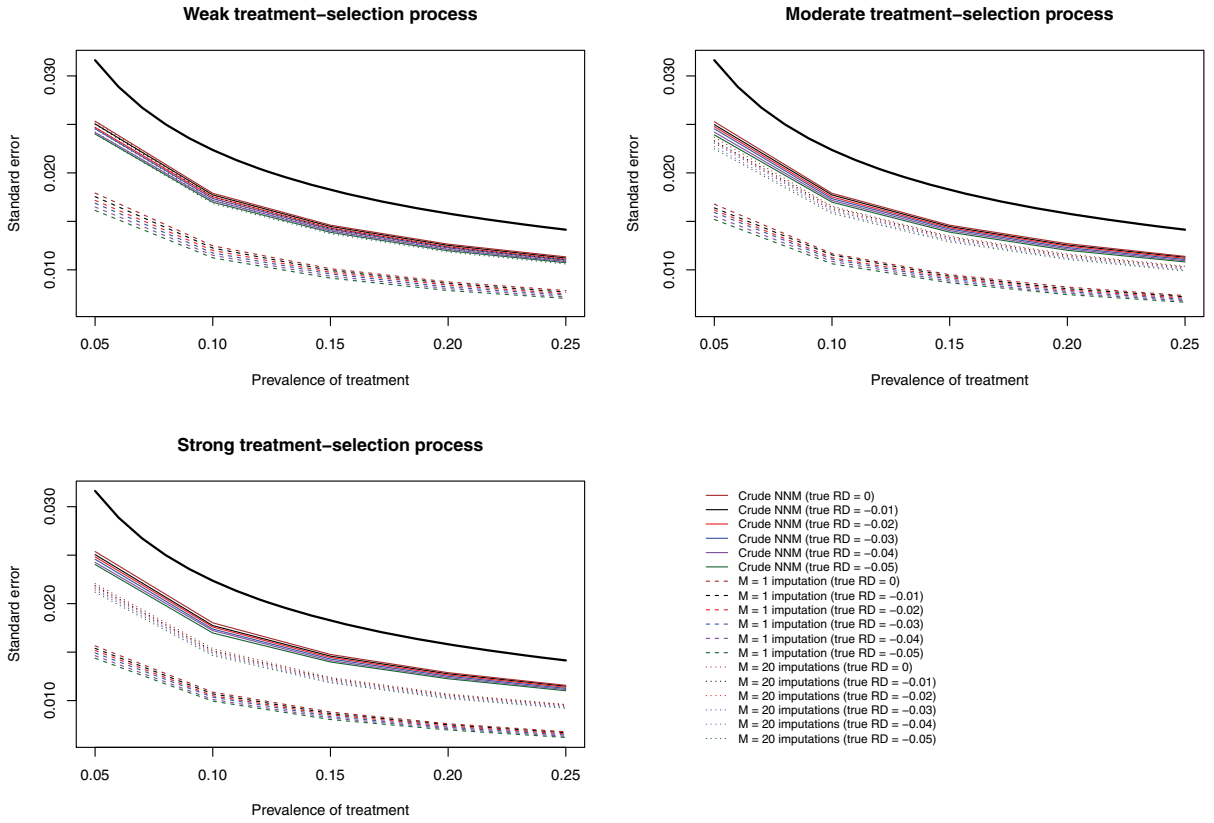

FIGURE 7 SE of the risk difference [Colour figure can be viewed at wileyonlinelibrary.com]

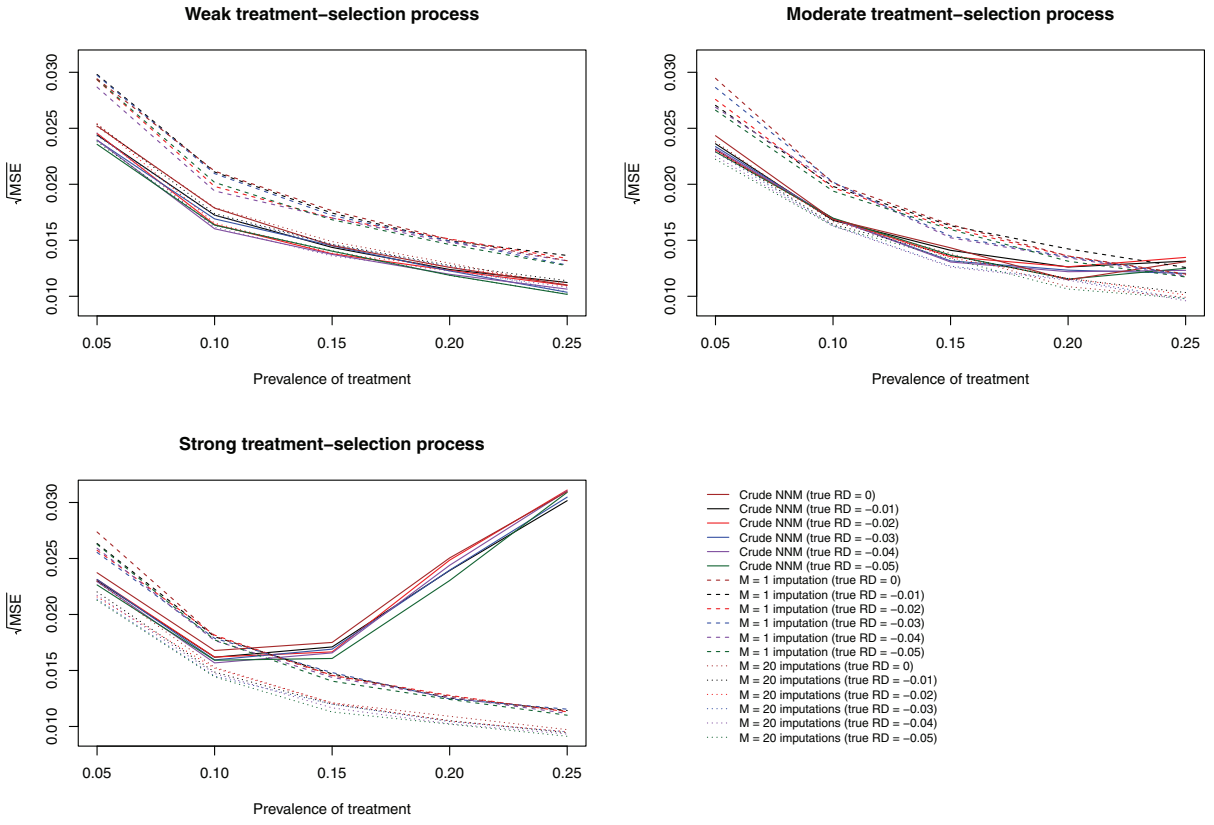

FIGURE 8 $\sqrt{\text{MSE}}$ of estimated risk difference [Colour figure can be viewed at wileyonlinelibrary.com]

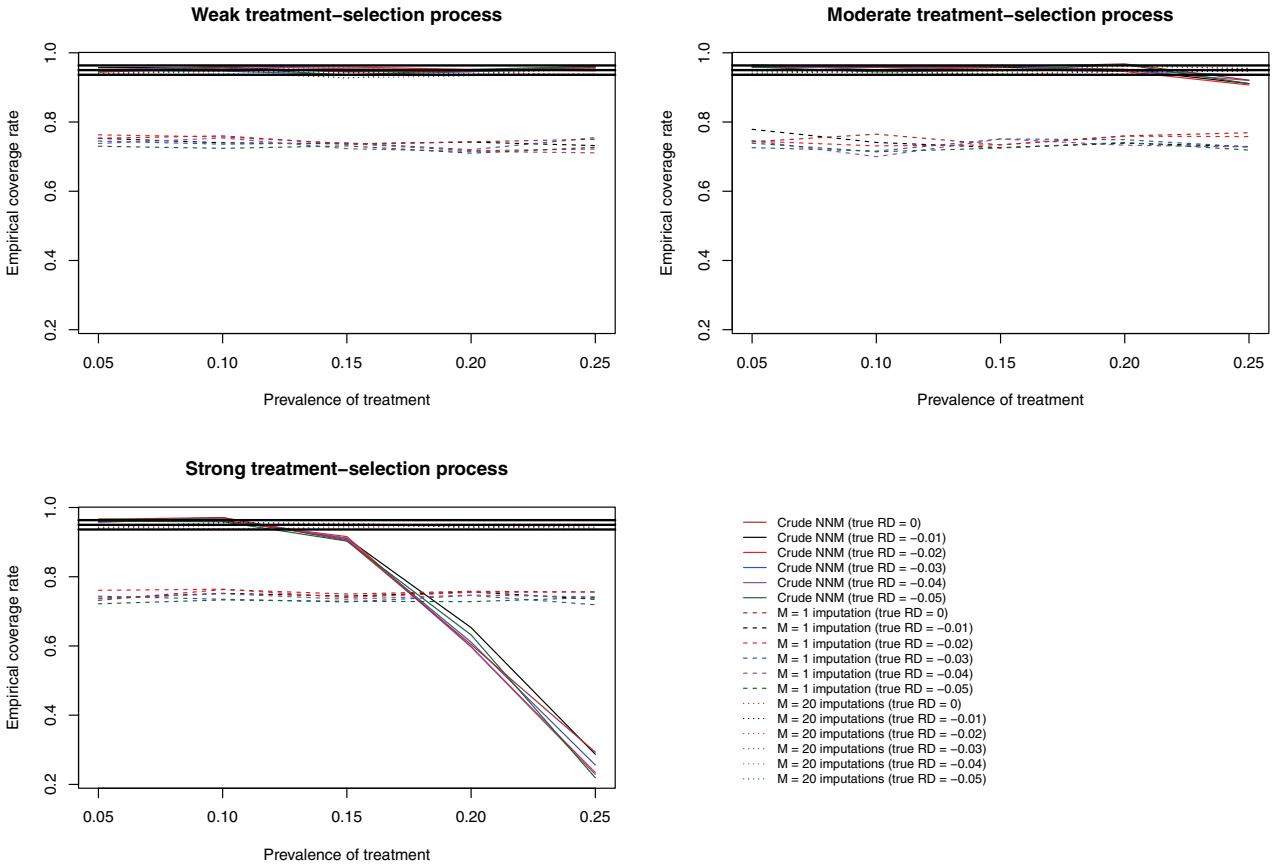

FIGURE 9 Empirical coverage rates of 95% confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

the treatment-selection process was strong, the SD of the sampling distribution was overestimated by at most approximately 10%. The use of single imputation of the potential outcomes under control resulted in an underestimation of the SD of the sampling distribution by approximately 40%. In contrast to this, the use of MI of potential outcomes under control (with $M = 20$), resulted in SEs that correctly approximated the SD of the sampling distribution of the risk difference.

4.8 | ANOVA of simulation results

We used analysis of variance (ANOVA) to examine the variation of the different metrics (bias, SD of the empirical sampling distribution of the estimated risk differences, estimated SE of the estimated risk difference, MSE, coverage of 95% confidence intervals, and ratio of SE to SD) across the following factors in the simulations: the true risk difference, the number of treated subjects (entered as $1/\sqrt{N}$), the number of imputed potential outcomes under control (entered as $1/\sqrt{M}$), and the strength of treatment selection process (entered as 1, 2, 3, for weak, medium, and strong). This analysis was restricted to scenarios with $M = 1, 5, 20$, and 100. Orthogonal polynomials were used to decompose the sum of squares for each factor. The residual error is estimated from the second and higher interactions of the factors, which were much smaller than the main effects. The resulting F-statistics are reported in Table 2. The simulated datasets are reused for each level of factor, M, which improves the comparisons between different numbers of imputed datasets, but also decreases the residual mean square error and thus somewhat inflates the reported F-statistics (however, we are not using these for statistical hypothesis testing, but for a quantitative assessment of the magnitude of the contribution of different factors).

As anticipated from theory and broad statistical experience, the precision of the model-based estimator, as measured by the empirical SD of the estimator, reported SE of the estimator, and root MSE, improves as a function of $1/\sqrt{(\text{sample size})}$. The bias in the model-based estimator also decreased with increasing sample size, but sample size explains much less of this small decrease. The decrease in bias with increasing matched sample size appears to be the result of the improved

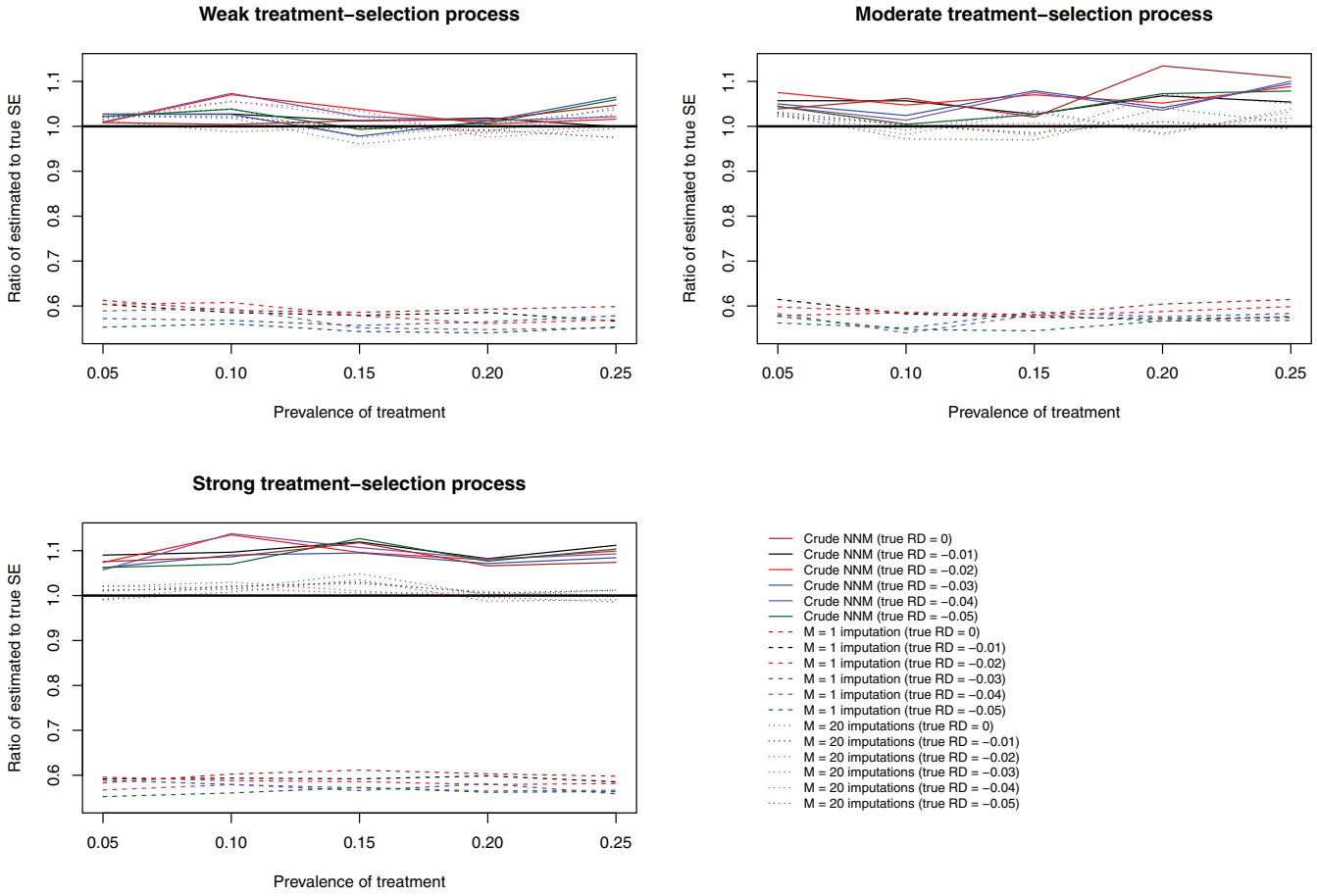

FIGURE 10 Ratio of mean SE to SD of sampling distribution [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 ANOVA of results of Monte Carlo simulations

Source	DF	Bias	SD	SE	MSE	Cover	SE/SD ratio
Number of treated subjects (N)							
Linear trend in $1/\sqrt{N}$	1	656.1	26 563.7	14 062	8237.6	0.1	13.3
Quadratic trend in $1/\sqrt{N}$	1	39.7	0	0.8	106.7	0	1.6
Deviation from linear and quadratic	2	10.1	0.5	0	0.4	4.2	3.8
True risk difference (RD)							
Linear trend in RD	1	52.1	63.7	52.6	24.3	13.1	7.1
Quadratic trend in RD	1	20.6	0.9	0	0.1	0.8	0.4
Deviation from linear and quadratic	2	1	0.1	0	0.1	1.2	0.2
Strength of treatment selection model							
Linear trend in strength	1	59.4	889.8	455.9	250.1	5.9	0.3
Quadratic trend in strength	1	3.5	0.2	0.2	0.1	5.2	0.3
Number of imputed outcomes under control (M)							
Linear trend in $1/\sqrt{M}$	1	0.4	1863.1	2517.5	524.7	27 496.8	28 174.1
Quadratic trend in $1/\sqrt{M}$	1	2.3	54.4	534.2	20.3	2580.8	3142.5
Deviation from linear and quadratic	1	0.7	0.3	14.7	0	20	39

Note: SD is the empirical SD of the estimated RD across simulation replicates; SE is the mean estimated SE of the estimated RD across simulation replicates; COVER is the empirical coverage rates of nominal 95% CIs; SE/SD RATIO is the ratio of mean estimated SE to the empirical SD of the estimated RD across simulation replicates.

estimation of the covariate-adjustment model. The simple difference in matched sample means displays a strong interaction between sample size and the strength of the covariates in the selection/outcome models, with increases in bias as the treated sample sizes increase and the resulting ratio of potential matches decreases.

Empirical coverage rates of 95% confidence intervals and the ratio of the mean estimated SE to the SD of the estimated risk difference across simulation replicates were both strongly related to the inverse of the square root of the number of imputed potential outcomes under control. The coverage of the intervals is generally good across all of the simulated conditions when $M = 20$.

4.9 | Sensitivity analyses—Nonlogistic link functions for treatment-selection model

In the supplemental online material, we report the results of a sensitivity analysis in which the link function for the treatment-selection model is misspecified. Results were comparable to those for when the link function was correctly specified.

5 | CASE STUDY

We provide a brief empirical example to illustrate the impact of the number of imputed datasets in an empirical analysis. The data consist of patients discharged from hospital with a diagnosis of congestive heart failure. The treatment is prescription of beta-blockers at hospital discharge.

5.1 | Data and analyses

We used the data that were described in Section 3.1. The propensity score was estimated by regressing receipt of a beta-blocker prescription on a set of 28 baseline covariates using a logistic regression model. The baseline covariates included demographic characteristics (age and sex), vital signs on admission (systolic blood pressure, respiratory rate, and heart rate), initial laboratory values (white blood count, hemoglobin, sodium, glucose, potassium, urea, and creatinine), comorbid conditions (diabetes, stroke or transient ischemic attack, previous AMI, atrial fibrillation, peripheral arterial disease, chronic obstructive pulmonary disease, dementia, cirrhosis, and cancer), presence of left bundle branch block on first electrocardiogram within 24 hours of admission, presenting signs and symptoms (neck vein distension, S3, S4, rales >50% of lung field), and findings on chest X-ray (pulmonary edema, cardiomegaly).

The methods in Section 2 were applied to these data. We examined the effect of the number of imputed datasets on the estimated risk difference for beta-blocker use by varying M from 1 to 100. For comparative purposes we also used conventional NNM in which the observed outcome of the matched control subject was used to replace the missing potential outcome under control for the matched treated subject. R code for conducting the analyses in the case study are available at the first author's GitHub repository [https://github.com/peter-austin/Stat_Med-2021-multiply-imputing-potential-outcomes-under-control-when-PS-matching].

5.2 | Results

In the original sample, 1895 (26.7%) of subjects received a prescription for a beta-blocker at hospital discharge. The Mahalanobis distance between the multivariate distribution of the baseline covariates in the treated and control subjects in the original sample was 0.316. The absolute value of the standardized differences for the baseline covariates ranged from 0.012 to 0.308 in the original sample. The probability of death within 1 year of hospital discharge was 0.229 and 0.315 in the treated and control subjects, respectively, corresponding to a crude risk difference of −0.086.

Using NNM, all treated subjects were matched to a control subject. The Mahalanobis distance between the multivariate distribution of the baseline covariates in the treated and control subjects was 0.008. The absolute value of the standardized differences for the baseline covariates ranged from 0.001 to 0.038. Thus, matching on the propensity score removed most of the differences between treated and control subjects. The probability of death within 1 year of hospital discharge was 0.229 and 0.283 in the matched treated and matched control subjects, respectively, for a risk difference

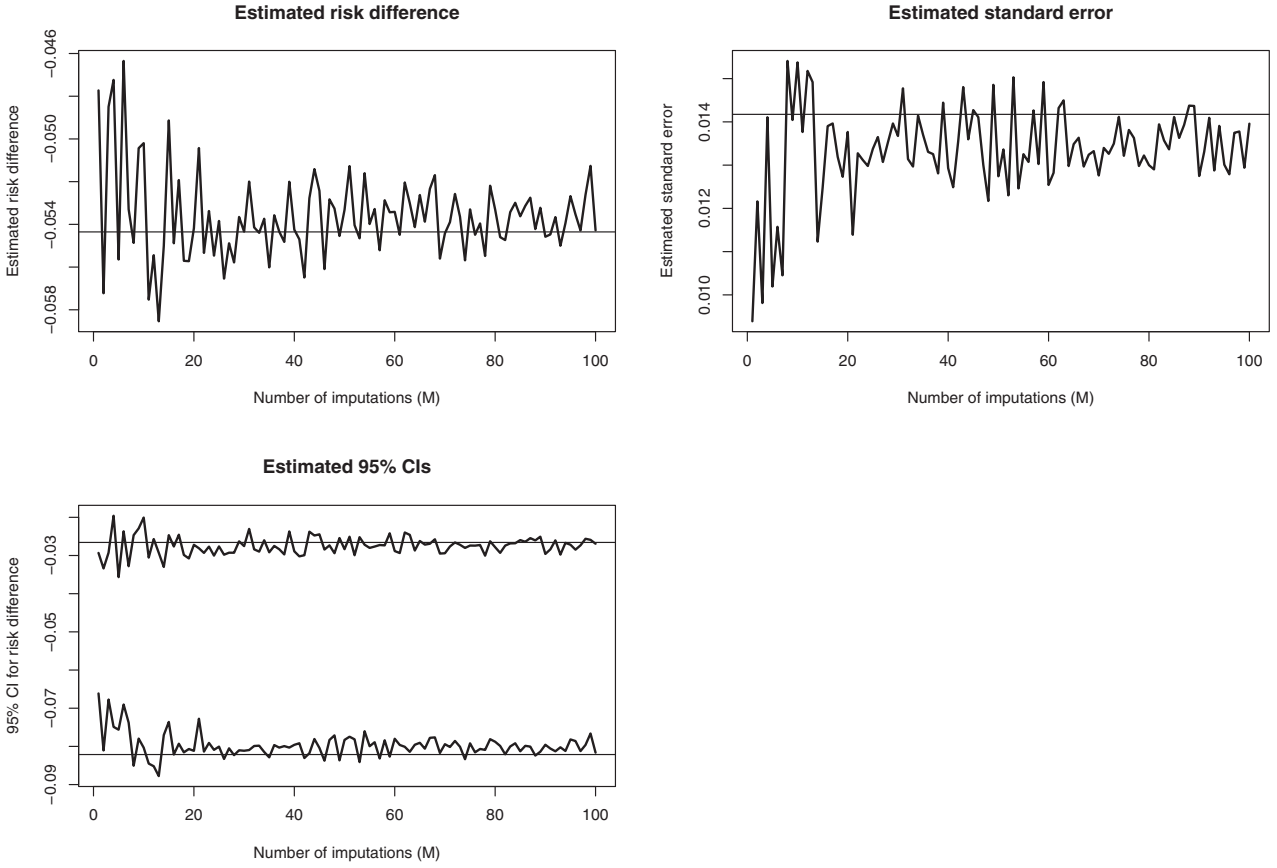

FIGURE 11 Analyses in case study

of −0.054 (95% CI: −0.082 to −0.027). The SE of the estimated risk difference was 0.014. This analysis represents the conventional analysis that would typically be done when using propensity-score matching.

The relationship between the number of imputed datasets and the estimated risk differences are described in the top left panel of Figure 11. The estimated risk differences ranged from −0.046 (when using $M = 6$) to −0.059 (when using $M = 13$). On this panel we have superimposed a horizontal line denoting the estimated risk difference of −0.054 obtained using the observed outcomes of the matched control subjects. The mean and median adjusted risk difference across the 100 values of M was −0.054, which agrees with the estimate obtained using crude NNM.

The relationship between the number of imputed datasets and the estimated SE of the estimated risk difference are reported in the top right panel of Figure 11. On this panel we have superimposed a horizontal line denoting the estimated SE when the observed outcomes for the matched control subjects were used. The estimated SE when using single imputation ($M = 1$) was, as expected, substantially lower than the SE obtained when using conventional NNM. Increases in M above 20 tended to result in negligible changes in the estimated SE. These results agree with those observed in the simulations.

The relationship between the number of imputed datasets and the estimated 95% confidence intervals are reported in the lower left panel. On this panel we have superimposed two horizontal lines denoting the estimated confidence interval when the observed outcomes for the matched control subjects were used. When using single imputation ($M = 1$), the estimated confidence intervals tended to be modestly narrower than when crude NNM or MI were used.

6 | DISCUSSION

We proposed a new method for making inferences about risk differences that combines matching on the propensity score with MI of potential outcomes under control for the matched treated subjects. The imputation of potential outcomes under control can be thought of as serving a function similar to that of postmatching regression adjustment. Having

proposed this method, we examined the effect of the number of imputed datasets on inferences about risk differences. We found that the additional covariate adjustment reduced the bias in the estimated risk difference compared with what was obtained when using conventional NNM. Imputing multiple imputed datasets resulted in estimated risk differences that were more efficient than using single imputation. The FMI reported in Section 4.2 is high (roughly 50%), consistent with the observation that increasing the number of imputations from the commonly suggested $M = 5$ to $M = 20$ resulted in further improvement in efficiency and performance of confidence intervals. This experience suggests using $M \geq 20$ in applications. There was no appreciable improvement in the simulated settings using M larger than 20, but the FMI should be checked to ensure it is not higher than for the settings evaluated here.

Earlier studies suggested imputing a single potential outcome under control for each matched treated subject using regression adjustment.[16,23] Our findings suggest that, although this approach reduces bias to the same extent as imputing multiple potential outcomes under control, it is statistically inefficient. Furthermore, it is an open question as to how to correctly estimate the SE of the resultant risk difference, as one does not account for between-imputation variation. In our simulations, when using single imputation, we acted as though the imputed data were known, and not estimated. Thus, the uncertainty implicit in the imputed quantity was ignored. This resulted in estimated SEs that were too small. Thus, our proposed method is an improvement over previously described methods that imputed a single potential outcome under control, as it is more statistically efficient and one can develop an appropriate variance estimator. Furthermore, the proposed method results in a greater reduction in bias than was observed for conventional NNM in the presence of a strong treatment-selection process and a low to moderate prevalence of treatment. Thus, the proposed method has superior performance to the existing methods of both conventional NNM and NNM combined with *single* imputation of potential outcomes under control.

We have described a variance estimator for our proposed estimator of the risk difference. Using Monte Carlo simulations, we showed that this variance estimator performed well across 90 scenarios defined by prevalence of treatment, true risk difference, and the strength of the treatment-selection process. The purpose of these simulations was to show that the variance estimator performed well across a range of realistic conditions provided that the propensity score model and the imputation model were both correctly specified. The performance of the MI-based variance estimator is not sensitive to the matching method used when the matching depends only on the baseline covariates and they are also included in the imputation model. Indeed, the variance estimator could be applied with the unmatched control sample. The matched control subjects are used only when fitting the outcome model, and inference about the fit is conditional on the covariates. Conditional on each set of model parameters simulated from the model fit, the "complete data" analysis correctly assumes the outcomes for different treated subjects are independent, consistent with routinely used modeling assumptions. Matching can reduce the variance of the treatment effect estimator by reducing the required model adjustment. This reduction is tracked by the variance estimator. However, the primary role of the matching is to reduce dependence on the specification of the imputation/outcome model, and to improve the approximate estimated model by restricting the estimation to the range of covariate values similar to those of treated subjects where it will be used.

Rosenbaum and Rubin coined the term "bias due to incomplete matching".[12] This refers to the bias that can occur when some treated subjects are excluded from the final matched sample. This bias arises because the average treatment effect in the treated is the target estimand when using matching. Excluding treated subjects can result in biased estimation if there are systematic differences between the matched treated subjects and the unmatched treated subjects. Although caliper matching can result in a greater reduction in bias due to confounding variables than does the use of NNM, it can also result in the exclusion of some treated subjects. An advantage to combining regression adjustment with matching is that one can use NNM, and thereby avoid bias due to incomplete matching, while simultaneously achieving greater reduction in bias due to confounding that would be possible from matching alone.

When outcomes are binary, there are four simple possible measures of treatment effect: the risk difference (also known as the absolute risk reduction), the relative risk, the number needed to treat (NNT—the reciprocal of the risk difference), and the odds ratio. Several clinical commentators have suggested that the odds ratio is of limited use for clinical decision making.[40-42] In the current study we have focused on estimation of the risk difference as it is a well-defined causal estimand within Rubin's Causal model. However, the methods described above can also be used to estimate relative risks. Thus, regression adjustment can be combined with propensity-score matching to estimate all of the quantities necessary to inform clinical decision making: the risk difference, the NNT, and the relative risk.

To the best of our knowledge, only one prior study has examined combining regression adjustment with a propensity score method to multiply impute potential outcomes under control. Gutman and Rubin proposed a method called multiple imputation with two splines and subclassification (MITSS).[19] There are three key differences between MITSS and the approach proposed in the current study. The first difference is that they considered subclassification (or stratification)

on the propensity score, whereas we used propensity-score matching. The former entails stratifying the entire sample into coarser strata based on the quintiles of the estimated propensity score. They considered the case in which there was a single continuous confounding variable (possibly the estimated propensity score) and subsequently regressed the outcome on this variable using spline functions. The fitted model was then used to multiply impute potential outcomes. The second difference is that they focused on estimating the marginal odds ratio, whereas we focused on estimating a causal risk difference. As noted above, our method can easily be modified to estimate the relative risk or the NNT. It could also be modified to estimate the marginal odds ratio. The third difference is that they only considered imputing 20 potential outcomes per subject, whereas we allowed this quantity to vary from 1 to 100 and examined the effect of this quantity on inferences about the risk difference. Although it is beyond the scope of the current study, our approach could be combined with MITSS to estimate adjusted risk differences when using stratification on the propensity score.

The primary limitation of the current study is its dependence on limited Monte Carlo simulations. However, we are not relying on the simulations to demonstrate the validity of the procedure; rather, they are used to assess its performance in a wide range of practical settings. Importantly, these settings reflected a range of true risk differences, prevalence of treatment, and strength of the treatment-selection process, and were informed by empirical analyses of the data that were subsequently used in our case study. It is possible that observed performance would vary under different data-generating processes and under different scenarios than our 90 scenarios. A second limitation of these simulations is that the regression model used for subsequent adjustment was identical to that used to generate outcomes. However, it is important to examine the performance of a statistical method when the models have been correctly specified. The performance of these methods when the regression model used for adjustment is misspecified merits exploration in future research.[43] We note that the assumption that the imputation model has been correctly specified is a common assumption when using MI in any context. A third limitation is that in the data-generating process the two potential outcomes were independent conditional on their covariates.

The objective of the current study was to describe a new method for estimating risk differences using observational data and to then demonstrate the performance of the proposed method. Our intention was not to provide a survey of methods for causal inference nor to compare the relative performance of different methods for causal inference. We refer the interested reader to articles describing different matching and weighting-based approaches to estimating causal effects.[3,5,44-51] A particular class of methods combine IPTW using the propensity score with regression adjustment.[46,47,49] This class of methods have been described as having a "double robustness" property, meaning that they provide consistent estimators of the average treatment effect if either the outcomes regression model or the propensity score model is specified correctly.[47] However, these methods may not have superior performance to single model strategies when both regression models are misspecified.[46]

Future research is necessary to compare the performance of the proposed method with other methods that combine model-based (eg, regression) adjustment with propensity score methods. In particular, such comparisons could include: MITSS (with an appropriate modification to estimate a marginal risk difference rather than a marginal odds ratio),[19] double propensity score adjustment,[22] augmented inverse propensity score weighting,[45] and Abadie and Imben's bias-corrected matching estimator.[21] Nonpropensity score-based approaches, such as targeted learning, also merit inclusion in such a comparison.[52] Due to space constraints, such comparisons are beyond the scope of the current study.

The proposed method requires the specification of two regression models: one for estimating the propensity score and one for imputing potential outcomes under control. In applied applications, relaxing the linearity assumption and allowing for continuous variables to have nonlinear relationships with the outcome (eg, through the use of restricted cubic splines), increases the likelihood that the models will be correctly specified.

In summary, we have described a new method that combines matching on the propensity score with MI of potential outcomes under control. The use of this method permits greater reduction in bias than is achieved with NNM alone. Using single imputation of adjusted potential outcomes under control, although achieving substantial reduction in bias compared with NNM alone, is less efficient than MI. Furthermore, it resulted in an estimated SE that underestimated the variance of the sampling distribution of the risk difference. In contrast to this, multiply imputing potential outcomes under control resulted in an MI estimate of the SE of the risk difference that accurately approximated the SD of the sampling distribution of the risk difference. Increasing the number of imputed potential outcomes under control resulted in more efficient estimation, with the SD of the sampling distribution of the estimated risk difference decreasing with increasing number of these imputed potential outcomes. The greatest relative increase in efficiency was achieved by imputing five potential outcomes under control for each of the matched treated subjects. Once 20 outcomes under control were imputed for each matched treated subject, further improvements in efficiency were negligible. We suggest that analysts impute 20

potential outcomes under control as this results in efficient estimation of risk differences and also results in a reduction in bias compared with the use of NNM alone.

ACKNOWLEDGEMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). This study also received funding from: Canadian Institutes of Health Research (CIHR) (PJT - 166161) (Dr. Austin). Dr. Austin is supported by a Mid-Career Investigator award from the Heart and Stroke Foundation. Professor Rubin is supported in part by grants from the US NIH, US NSF, and the US ONR. Parts of this material are based on data and information compiled and provided by: MOH. The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred.

DATA AVAILABILITY STATEMENT

The dataset from this study is held securely in coded form at ICES. While data sharing agreements prohibit ICES from making the dataset 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
Neal Thomas https://orcid.org/0000-0002-1915-8487

REFERENCES

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
2. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. *Multivar Behav Res*. 2011;46:119-151.
3. Austin PC. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46:399-424.
4. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516-524.
5. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866.
6. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*. 2004;13(12):841-853.
7. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037-2049.
8. 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.
9. 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.
10. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39:33-38.
11. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med*. 2014;33(6):1057-1069.
12. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics*. 1985;41(1):103-116.
13. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*. 1973;29:185-203.
14. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J Am Stat Assoc*. 1979;74(366a):318-328.
15. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J Am Stat Assoc*. 2000;95:573-585.
16. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86:4-29.
17. Rubin DB. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Cambridge, Massachusetts: Harvard University; 1970.
18. Belson WA. A technique for studying the effects of a television broadcast. *J R Stat Soc Ser C Appl Stat*. 1956;5(3):195-202.
19. Gutman R, Rubin DB. Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Stat Med*. 2013;32(11):1795-1814.
20. Quade D. Nonparametric analysis of covariance by matching. *Biometrics*. 1982;38(3):597-611.
21. Abadie A, Imbens GW. Bias-corrected matching estimators for average treatment effects. *J Bus Econ Stat*. 2011;29(1):1-11.

22. Austin PC. Double propensity-score adjustment: a solution to design bias or bias due to incomplete matching. *Stat Methods Med Res*. 2017;26(1):201-222.
23. 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(3):728-751.
24. Austin PC, Manca A, Zwarenstein M, Juurlink DN, Stanbrook MB. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *J Clin Epidemiol*. 2010;63(2):142-153.
25. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.
26. van Buuren S. *Flexible Imputation of Missing Data*. 2nd ed. Boca Raton, FL: CRC Press; 2018.
27. Carpenter JR, Kenward MG. *Multiple Imputation and its Application*. Chichester, UK: John Wiley & Sons; 2013.
28. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Chichester, UK: John Wiley & Sons; 2007.
29. Molenberghs G, Fitzmaurice G, Kenward MG, Tsiatis A, Verbeke G. In: Fitzmaurice G, ed. *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC; 2015 Handbook of Modern Statistical Methods.
30. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med*. 1988;318:1728-1733.
31. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *Br Med J*. 1995;310(6977):452-454.
32. Rubin DB. Bayesian inference for causality: the importance of randomization. Paper presented at: Proceedings of the Social Statistics Section; American Statistical Association; 1975;233-239.
33. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6:34-58.
34. 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.
35. Agresti A, Min Y. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat Med*. 2004;23(1):65-75.
36. Meng XL. Multiple-imputation inferences with uncongenial sources of input. *Stat Sci*. 1994;9(4):538-573.
37. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making*. 2009;29(6):661-677.
38. 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.
39. Austin PC. Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Commun Stat Simul Comput*. 2009;38:1228-1234.
40. Sackett DL. Down with odds ratios! *Evid Based Med*. 1996;1:164-166.
41. Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *Can Med Assoc J*. 1995;152(3):351-357.
42. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol*. 1994;47(8):881-889.
43. 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.
44. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1-21.
45. Glynn AN, Quinn KM. An introduction to the augmented inverse propensity weighted estimator. *Polit Anal*. 2010;18:36-56.
46. Kang J, Schafer J. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22:523-580.
47. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13(4):279-313.
48. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387-394.
49. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models: review and new applications. *Am Stat*. 2004;58:272-279.
50. Curtis LH, Hammill BG, Eisenstein EL, Kramer JM, Anstrom KJ. Using inverse probabilty-weighted estimators in comparative effectiveness analyses with observational databases. *Med Care*. 2007;45(10 Supplement 2):S103-S107.
51. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960.
52. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer; 2011.

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, Rubin DB, Thomas N. Estimating adjusted risk differences by multiply-imputing missing control binary potential outcomes following propensity score-matching. *Statistics in Medicine*. 2021;40(25):5565-5586. https://doi.org/10.1002/sim.9141