STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:1402–1414
Published online 16 February 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3549

Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach

Yongming Qu*,† and Ilya Lipkovich

Eli Lilly and Company, Indianapolis, IN 46285, U.S.A.

SUMMARY

Propensity scores have been used widely as a bias reduction method to estimate the treatment effect in nonrandomized studies. Since many covariates are generally included in the model for estimating the propensity scores, the proportion of subjects with at least one missing covariate could be large. While many methods have been proposed for propensity score-based estimation in the presence of missing covariates, little has been published comparing the performance of these methods. In this article we propose a novel method called multiple imputation missingness pattern (MIMP) and compare it with the naive estimator (ignoring propensity score) and three commonly used methods of handling missing covariates in propensity score-based estimation (separate estimation of propensity scores within each pattern of missing data, multiple imputation and discarding missing data) under different mechanisms of missing data and degree of correlation among covariates. Simulation shows that all adjusted estimators are much less biased than the naive estimator. Under certain conditions MIMP provides benefits (smaller bias and mean-squared error) compared with existing alternatives. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: propensity score; multiple imputation; missingness pattern; multiple imputation missingness pattern; inverse probability weighted estimator

1. INTRODUCTION

Observational studies are becoming increasing important, as they allow us to observe treatment outcomes for a large number of subjects in 'real-world' treatment practice. Well-designed observational studies provide valuable information in addition to randomized controlled trials [1]. Evaluating the cause–effect relationship from observational studies is a compelling statistical problem. A challenge for analyzing observational studies is how to estimate the treatment effect when treatment is not randomly assigned. The propensity score method introduced by Rosenbaum and Rubin [2] is a tool for estimating a causal effect of a nonrandomized treatment in the presence of

*Correspondence to: Yongming Qu, Lilly Research Laboratory, Eli Lilly and Company, Indianapolis, IN 46285, U.S.A.
†E-mail: qu_yongming@lilly.com

Received 23 December 2007
Accepted 23 December 2008

Table I. Illustration of the method of fitting separate regression models for each missingness pattern for estimation of the propensity score when there are two covariates.

Index	Missingness pattern	Variables to be included	# Observations
1	X_1 is missing; X_2 is missing	Intercept	n_1
2	X_1 is missing; X_2 is nonmissing	Intercept, X_2	n_2
3	X_1 is nonmissing; X_2 is missing	Intercept, X_1	n_3
4	X_1 is nonmissing; X_2 is nonmissing	Intercept, X_1, X_2	n_4

imbalance of baseline covariates (**X**) between treatment groups. The propensity score, which is the probability of a subject to be assigned to a certain treatment given **X**, is essentially a mapping of multiple covariates onto a one-dimensional variate. Propensity scores are typically estimated via a multiple logistic regression.

It has been shown that using propensity scores results in substantial reduction of bias in estimating the treatment effect when treatment assignments are subject to selection bias [2]. Furthermore, the propensity score method provides advantages compared with simply incorporating all the covariates in the model for treatment effect. For example, the propensity score method is more robust compared with direct covariate adjustment with respect to model over-parameterization and unequal covariance matrices for treated and untreated groups (see Reference [3, p. 2278]).

As the propensity score method uses many covariates and there are generally a lot of missing values in observational studies, a large proportion (e.g. 30–50 per cent) of subjects could have missing values for at least one covariate. Therefore, ignoring the missing values could result in deleting a large number of observations. A simple and intuitive way for handling categorical missing values is to create a new category for the variable with missing values to indicate missingness. However, this ignores the correlation among covariates and therefore is not an efficient method. A more sophisticated method is to fit separate regressions in estimation of the propensity score for each distinct missingness pattern [4]. See Table I for an illustration of this method.

Although this method includes all nonmissing values for those subjects with the same missingness pattern, it increases the variability in the estimated propensity scores since only a subset of subjects are included in each propensity score model. A much more complicated and computational intensive approach is to jointly model the propensity score and the missingness, and then using EM/ECM algorithm or Gibbs sampling to estimate the parameters and propensity score [5–7]. Since there is no package in the standard statistical software (e.g. SAS, R or S-PLUS) to perform such analysis, the computational complexity makes this approach less attractive and less practical.

As a different approach, the propensity score with missing values could also be estimated using the general methodology of multiple imputation proposed by Rubin [8, 9]. Recently, Crowe *et al.* [10] studied the performance of the estimation of propensity scores using multiple imputation through simulation. The idea of multiple imputation is to fill the missing values multiple times through sampling from the posterior predictive distribution of the missing values given the observed values. One advantage is that the data from each imputed sample can be analyzed using 'standard methods'. Another advantage is that the multiple imputation procedure allows us to include ancillary variables, which may not directly affect the propensity scores. The purpose of multiple imputation (compared with single conditional mean imputation) is to give not only the point estimator of a parameter but also the variance for the point estimator. In applying multiple imputation to the estimation of the treatment effect by the propensity score method, it should be noted that the

DOI: 10.1002/sim

usual multiple imputation variance estimator ('Rubin's rule') may not work because it does not account for the uncertainty in propensity scores estimated by the logistic regression, which is generally difficult to quantify by a closed-form expression. However, the point estimator for the treatment difference by multiple imputation remains valid. The variance in the point estimate could be evaluated by jackknife or bootstrap methods. Surprisingly, there is little research in applying multiple imputation in estimating the propensity scores in the presence of missing values, nor research in comparing various methods through simulation.

Here, we will introduce a new method called *multiple imputation missingness pattern* (MIMP), which utilizes the multiple imputation method and the pattern of missing data in the estimation of propensity scores. The remainder of this article is organized as follows. Section 2 introduces notation and outlines the MIMP method. In Section 3, we compare the MIMP method with existing methods through simulation. In Section 4, we apply various methods to a set of data from a clinical study. Finally, Section 5 includes summary and discussion.

2. METHODS

In this section, we consider two treatment groups: $T=0$ for the reference group and $T=1$ for the experimental treatment group. Let T be distributed as a Bernoulli random variable with parameters defined from the following logistic regression model:

$$\text{logit}(\Pr(T_j=1|\mathbf{X}_j,\boldsymbol{\beta}))=(1,\mathbf{X}_j')\boldsymbol{\beta},\quad j=1,2,\ldots,n \tag{1}$$

where $\mathbf{X}_j=(X_{1j},X_{2j},\ldots,X_{kj})'$, $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_k)$, k is the number of covariates, and n is the number of subjects (i.e. sample size). If there are no missing values, the propensity scores are estimated by

$$\hat{p}_j=\frac{\exp\{(1,\mathbf{X}_j')\hat{\boldsymbol{\beta}}\}}{1+\exp\{(1,\mathbf{X}_j')\hat{\boldsymbol{\beta}}\}},\quad j=1,2,\ldots,n \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is a vector of the estimated regression coefficients from the logistic regression (1).

There are several ways of using propensity scores to estimate the treatment effect (see Reference [3] for a comparative review). In this article, we use the inverse-probability weighted (IPW) estimator [11–13] to estimate the treatment effect (θ_d):

$$\hat{\theta}_d=\hat{\theta}_1-\hat{\theta}_0 \tag{3}$$

where

$$\hat{\theta}_0=\sum_{j=1}^{n}[1-\hat{p}_j]^{-1}Y_j(1-T_j) \tag{4}$$

$$\hat{\theta}_1=\sum_{j=1}^{n}[\hat{p}_j]^{-1}Y_jT_j \tag{5}$$

and $Y_1,Y_2,\ldots,Y_n$ are the responses for subject $1,2,\ldots,n$, respectively.

Although the propensity score stratification is more popular than the IPW estimator in current medical research, it was shown via theoretical argument and simulation [14] that IPW is less biased and more efficient than the propensity score stratified estimators, and IPW also attains

DOI: 10.1002/sim

the semi-parametric efficiency guaranteed by theory introduced in Robins *et al.* [15]. In case that the propensity score model is correct, the IPW approach gives a consistent estimator while other methods merely reduce bias. Therefore, IPW is a more sensitive estimator to detect the performance difference in the simulation. Although the IPW estimator is used exclusively in the article, the methods for propensity score estimation with missing values can certainly be used in combination with any propensity score-based methods (e.g. subclassification) for estimating the treatment effect.

In the presence of missing values, the multiple imputation method (referred to as the MI method hereafter) or the method of estimating the propensity scores separately for distinct missingness patterns (referred to as the MP method hereafter) could be used. For a brief overview of the idea of multiple imputation, we assume $\mathbf{X}=(X_1, X_2, \ldots, X_k)'$ are distributed from a multivariate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, for subject j, the distribution of the unobserved values ($\mathbf{X}_j^{(u)}$) given the observed values ($\mathbf{X}_j^{(o)}$) is also a normal distribution. Instead of imputing the missing value with the conditional mean, multiple imputation imputes the missing value multiple times with values which are generated from the conditional distribution. The advantage of multiple imputation is that it automatically considers the variability in the imputed value. In practice, the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown; hence, Gibbs samples are required to generate multiple imputed samples from the posterior predictive distribution of the missing data given the observed data. Little and Rubin [16] classified the mechanisms of missingness into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Multiple imputation generally works for MCAR and MAR, but ignores the missingness after imputing the data, which was originally viewed as an advantage because 'standard methods' could be applied. As a result, it does not fully utilize the information provided by missingness itself which may help to better estimate the propensity scores under MNAR. On the other hand, the MP method utilizes the information provided by missingness itself, but it increases the variability in the estimated propensity scores when fitting separately models for individual missingness patterns. To see why missingness itself could provide additional information for the missing values, we assume an extreme case where one covariate X_1 can have value of 0 or 1, and X_1 is missing when $X_1=1$. In this case, multiple imputation could not impute the missing value for X_1 well as the missingness is not at random. However, the missing indicator variable for X_1 provides perfect information for X_1. Of course, in reality this is rarely the case, but missingness itself could provide partial information for the missing values in addition to the observed values of $\mathbf{X}$. To overcome the drawbacks of the MI and MP methods, we propose a new method called *MIMP*. The idea of MIMP is simple. First, impute the missing values using multiple imputation. Second, estimate the propensity scores by a logistic regression with the imputed values of $\mathbf{X}$ and the missingness pattern variable (say, S) as independent variables. The procedure of MIMP is outlined below:

1. Impute the data with multiple imputation using $\mathbf{T}$, $\mathbf{Y}$ and the nonmissing values of $\mathbf{X}$, where $\mathbf{T}=(T_1, T_2, \ldots, T_n)'$, $\mathbf{Y}=(Y_1, Y_2, \ldots, Y_n)'$ and $\mathbf{X}=(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)'$. Let $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(M)}$ denote M imputed samples of $\mathbf{X}$.
2. For each imputed sample $\mathbf{X}^{(m)}$, the propensity scores $\hat{p}_1^{(m)}, \hat{p}_2^{(m)}, \ldots, \hat{p}_n^{(m)}$ are estimated from a linear logistic regression model with independent variables $\mathbf{X}^{(m)}$ and S.
3. The treatment difference from the mth imputed sample could be estimated by

$$\hat{\theta}_d^{(m)} = \hat{\theta}_1^{(m)} - \hat{\theta}_0^{(m)} \tag{6}$$

DOI: 10.1002/sim

where

$$\hat{\theta}_0^{(m)} = \sum_{j=1}^{n} [1-\hat{p}_j^{(m)}]^{-1} Y_j (1-T_j) \tag{7}$$

and

$$\hat{\theta}_1^{(m)} = \sum_{j=1}^{n} [\hat{p}_j^{(m)}]^{-1} Y_j T_j \tag{8}$$

4. Combining the estimates from multiple imputed data sets, we have

$$\hat{\theta}_d = M^{-1} \sum_{m=1}^{M} \hat{\theta}_d^{(m)}, \quad \hat{\theta}_0 = M^{-1} \sum_{m=1}^{M} \hat{\theta}_0^{(m)} \quad \text{and} \quad \hat{\theta}_1 = M^{-1} \sum_{m=1}^{M} \hat{\theta}_1^{(m)} \tag{9}$$

The MI method is similar to the above procedure except in step (2), the propensity scores are estimated from a linear logistic regression model with independent variables $\mathbf{X}^{(m)}$, whereas in MIMP the regression model also includes indicator variables (S), capturing information about the missingness pattern.

There is a practical problem for methods utilizing the missingness pattern information (such as MIMP or MP). The number of subjects in some patterns may not be large enough to produce stable estimators for propensity scores. To address this problem, one could pool patterns of missing data with small number of observations. Here is an algorithm for combining small patterns to ensure that each pooled cell has at least $d_{\min}$ observation. Let $S=1,2,\ldots,B$ denote B patterns ordered from the largest pattern to the smallest pattern, and $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_B$ denote the corresponding numbers of observations. If $d_1 \geqslant d_{\min}$, the first pattern already has enough observations; hence, it remains the same. Otherwise, combine the first pattern with the most similar pattern among all those patterns having the number of observations less than $d_{\min}$ (if there are several equally similar candidate patterns, we choose the one with the smallest number of observations) until the number in the combined cell reaches $d_{\min}$. The dissimilarity between two patterns is defined as the Euclidean distance of the corresponding indicator vectors whose elements are set to 1 for missing and 0 for nonmissing. For example, if there are three variables, we use vector $(1,0,0)'$ to indicate the missingness pattern of (X_1 = missing, X_2 = nonmissing, X_3 = nonmissing). Repeat the above steps for the rest of the patterns. A pattern that has no less than $d_{\min}$ observations or has been pooled is said to be 'processed'. In the end of this process, there may be a situation when the total number of observations in the 'unprocessed' patterns is less than $d_{\min}$. Then, each of those patterns will be combined with the most similar among 'processed' patterns.

We illustrate this algorithm to pool patterns exceeding $d_{\min}=100$ for a situation where there are 500 observations with eight distinct missingness patterns formed by three covariates (Table II):

1. Since pattern 1 has more than $d_{\min}=100$ observations, there is no need to pool this pattern.
2. Since pattern 2 has more than $d_{\min}=100$ observations, there is no need to pool this pattern.
3. Pattern 3 has only 90 observations. The distances between pattern 3 and patterns 4–8 are $\sqrt{2}$, 1, $\sqrt{2}$, $\sqrt{2}$ and 1, respectively. Patterns 5 and 8 have the shortest distance from pattern 3, but pattern 8 is smaller than pattern 5. Therefore, we combine pattern 3 with pattern 8 to have a total of 110 observations.
4. Pattern 4 has only 70 observations. The distances between pattern 4 and patterns 5–7 are $\sqrt{3}$, $\sqrt{2}$ and $\sqrt{2}$, respectively. Since pattern 7 is smaller than pattern 6, we pool patterns 4

DOI: 10.1002/sim

Table II. Number of observations in each missingness pattern.

Index	Missingness pattern (S)	Number of observations (d)	Pooled cell
1	(0, 0, 0)	120	(1)
2	(1, 0, 1)	100	(2)
3	(0, 1, 0)	90	(3)
4	(0, 0, 1)	70	(4)
5	(1, 1, 0)	45	(3)
6	(1, 0, 0)	30	(4)
7	(1, 1, 1)	25	(4)
8	(0, 1, 1)	20	(3)

and 7 to have 95 observations, which is still smaller than 100. Then, we continue combining pattern 6 with the newly formed cell to have 125 observations.

5. Now only pattern 5 with 45 observation is left. Patterns 3, 6 and 7 have the shortest distance from pattern 5. Since pattern 3 has the largest number of observations, we pool pattern 5 with the pooled cell 3 in which pattern 3 belongs, to have 155 observations. The reason we combine pattern 5 with the most similar pattern with the largest number of observation is to maximize the similarity (i.e. maximize the number of observations that have the smallest distance from the candidate pattern). There are other ways to achieve the 'optimal' similarity, but it will have little effect on the estimation of the propensity score since there are only a small number of observations (less than $d_{\min}$) that are affected.

Finally, there are four new patterns with 120, 100, 155 and 125 observations. The above algorithm guarantees that each of the newly pooled cells has a minimum number of observations $d_{\min}$ and the patterns are similar within each cell.

On the basis of Rubin's formula [9], the variance in the MIMP estimator is expressed as

$$V_t = M^{-1} \sum_{m=1}^{M} V_m + (M+1) M^{-1} V_b \tag{10}$$

where V_m is the variance of $\theta_d^{(m)}$ given the imputed sample and V_b is the between imputation variance estimated by

$$\hat{V}_b = (M-1)^{-1} \sum_{m=1}^{M} (\hat{\theta}_d^{(m)} - \hat{\theta}_d)^2 \tag{11}$$

In practice, the variance V_m is not easy to estimate because there is additional uncertainty due to the estimated weights in the IPW estimator. Therefore, the variance of the MIMP estimator cannot be estimated directly from Rubin's formula. Tu and Zhou [17] gave a two-sample bias-corrected accelerated (BCa) bootstrap method for the estimator of the treatment difference based on the propensity score approach. Tu and Zhou argued that the treatment difference estimation is a two-sample problem; hence, a special BCa estimation procedure should be used. However, they assumed that the two estimators $\hat{\theta}_0$ and $\hat{\theta}_1$ are independent. As shown in (2), the propensity scores used to estimate $\hat{\theta}_0$ and $\hat{\theta}_1$ are based on the same $\hat{\boldsymbol{\beta}}$, which is estimated using the same model

DOI: 10.1002/sim

from the same subjects, hence, $\hat{\theta}_0$ and $\hat{\theta}_1$ are correlated. Actually, this is essentially a one-sample problem because the treatment difference in (3) can be re-written as

$$\hat{\theta}_d = \sum_{j=1}^{n} \{[\hat{p}_j^{(m)}]^{-1} T_j - [1-\hat{p}_j^{(m)}]^{-1}(1-T_j)\} Y_j \tag{12}$$

Therefore, the usual BCa bootstrap method [18] could be used.

3. SIMULATION

In this section, we compare five methods through simulation: the unadjusted estimator (referred to as the NAIVE estimator), the estimator only including subjects with complete covariates (referred to as the COMP method), the MI method, the MP method and the MIMP method. The NAIVE estimator is simply the difference of the means for the two treatment groups, and for the other four methods, the IPW estimator is used to estimate the treatment difference. More specifically, for the COMP and MP methods, after the propensity scores are estimated, the treatment effect is estimated by (3). The MIMP method is based on the procedure outlined in Section 2. The MI method is similar to the MIMP method except that the missingness pattern variable is not included in the propensity score estimation model. Simulation is performed using R 2.4.1 and the multiple imputation is implemented using Multivariate Imputation by Chained Equations package with Bayesian linear regression [19, 20]. For each data set, $M=5$ imputed samples are generated.

The indicator for treatment assignment (T) is generated as a Bernoulli random variable with probability defined with

$$\text{logit}(\Pr(T_j=1|\mathbf{X}_j)) = (1, \mathbf{X}_j')\boldsymbol{\beta}, \quad j=1,2,\ldots,n \tag{13}$$

where $\mathbf{X}_j=(X_{1j}, X_{2j}, \ldots, X_{kj})'$, $\boldsymbol{\beta}=(\beta_0, \beta_1, \ldots, \beta_k)$, k is the number of covariates and n is the number of subjects (i.e. sample size).

The indicators of missing values for k covariates are also generated as Bernoulli random variables whose parameters are defined by a set of k logistic models

$$\text{logit}(\Pr(R_{lj}=1|\mathbf{X}_j)) = (1, \mathbf{X}_j')\boldsymbol{\alpha}_l, \quad j=1,2,\ldots,n, \quad l=1,2,\ldots,k \tag{14}$$

where $\boldsymbol{\alpha}_l=(\alpha_{l0}, \alpha_{l1}, \ldots, \alpha_{lk})'$, $R_{lj}=1$ if X_{lj} is missing, and $R_{lj}=0$ otherwise. The response variable Y is generated from a linear model

$$Y_j = (1, \mathbf{X}_j', T_j)\boldsymbol{\gamma} + \varepsilon_j \tag{15}$$

where $\boldsymbol{\gamma}=(\gamma_0, \gamma_1, \ldots, \gamma_k, \gamma_t)$ and ε_j's are identically independently distributed from a normal distribution $\text{N}(0, \sigma^2)$.

In the simulation, we choose $\sigma=3.0$, $k=12$, $\boldsymbol{\beta}=(0, 0.2, \ldots, 0.2)'$ and $\boldsymbol{\gamma}=(0, 0.2, \ldots, 0.2, 0, 0, 0, 1.0)'$, which gives a true treatment effect of 1.0. The covariates $\mathbf{X}_j$'s are generated independently from a multivariate normal distribution $\text{N}(\mathbf{0}, \Sigma)$, where the diagonal elements of Σ are equal to 1 and the off-diagonal elements are set to r. In the simulation, we choose r = 0, 0.3, 0.5 and 0.8 to model the correlation among the covariates. We compare the five methods for three simulation settings with three types of missingness: MCAR, MAR and MNAR. In simulation setting A, we choose $\boldsymbol{\alpha}_1=\cdots=\boldsymbol{\alpha}_6=(-0.62, 0, \ldots, 0)'$. That is, each of $X_1, X_2, \ldots, X_6$ could be missing with a

DOI: 10.1002/sim

probability of 0.35, and $X_7, \ldots, X_{12}$ are not missing for any subject. In this setting, the missingness is MCAR because the probability of missingness only depends on the intercept in model (14). In simulation setting B, we choose $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_6 = (-1.0, 0, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)'$. The probability for each of $X_1, X_2, \ldots, X_6$ to be missing depends only on $X_7, \ldots, X_{12}$ (which are observed for all subjects) through model (14). The missingness in this setting is MAR. In simulation setting C, we choose $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_{12} = (-2.0, 0.3, \ldots, 0.3)'$. It means that for each of the variables $X_1, X_2, \ldots, X_{12}$, the probability of a missing value depends on the observed and unobserved values of $X_1, X_2, \ldots, X_{12}$ through model (14). Therefore, the missingness for this setting is MNAR. In all settings, the proportion of missing values for each covariate (i.e. X_1–X_6 for setting A and B, and X_1–X_{12} for setting C) is approximately 20–35 per cent. Approximately 50 per cent of subjects were assigned to each treatment group.

To ensure the stability of the estimators for MIMP, we used the algorithm in Section 2 with $d_{\min} = 50$ to pool small patterns of missing data. We also suggest a slightly modified version of the MP method in our simulation and analysis. For estimation of the propensity score within a pooled cell for the MP method, we include the largest set of nonmissing variables with missing values imputed by the unconditional means. For example, assume a pooled cell includes three patterns: only X_1 is nonmissing, only X_2 is nonmissing and only X_3 is nonmissing. Then, the regression model used for estimating the propensity score for this pooled cell includes variables X_1–X_3 and the missing values of X_1–X_3 are imputed by the corresponding unconditional means.

Table III shows the simulation results for all settings with sample size $n = 500$. In all settings, the NAIVE estimator is severely biased, which is consistent with previous research demonstrating

Table III. Empirical bias, standard deviation (SD) and mean-squared error (MSE) from 5000 simulations ($n = 500$).

		Setting A			Setting B			Setting C		
r	Method	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
0	NAIVE	0.3257	0.2785	0.1836	0.3195	0.2680	0.1739	0.3266	0.2745	0.1820
	COMP	0.0125	0.6262	0.3923	0.0083	0.6719	0.4515	0.0123	0.6063	0.3677
	MP	0.0016	0.3161	0.0999	−0.0025	0.3240	0.1049	0.0136	0.3317	0.1102
	MI	0.0001	0.2991	0.0895	−0.0036	0.2979	0.0888	0.0201	0.3004	0.0906
	MIMP	−0.0001	0.2993	0.0895	−0.0036	0.2984	0.0890	0.0090	0.3013	0.0909
0.3	NAIVE	1.1118	0.2803	1.3147	1.1131	0.2796	1.3171	1.1203	0.2779	1.3323
	COMP	0.0601	0.8611	0.7449	0.0298	0.8492	0.7218	0.0190	0.7827	0.6128
	MP	0.0240	0.4521	0.2049	0.0211	0.4753	0.2263	0.0896	0.4599	0.2195
	MI	0.0134	0.3943	0.1557	0.0167	0.3988	0.1593	0.1112	0.3894	0.1639
	MIMP	0.0138	0.3955	0.1565	0.0165	0.3997	0.1600	0.0755	0.3772	0.1479
0.5	NAIVE	1.5152	0.2853	2.3772	1.5218	0.2877	2.3986	1.5136	0.2802	2.3694
	COMP	0.1291	1.0049	1.0263	0.0742	1.0263	1.0586	0.0762	0.8639	0.7520
	MP	0.0521	0.5962	0.3581	0.0533	0.6111	0.3762	0.1629	0.5745	0.3565
	MI	0.0281	0.5250	0.2763	0.0415	0.5002	0.2519	0.2011	0.4598	0.2518
	MIMP	0.0289	0.5271	0.2787	0.0412	0.5007	0.2523	0.1513	0.4209	0.2001
0.8	NAIVE	2.0275	0.2904	4.1951	2.0245	0.2904	4.1831	2.0308	0.2920	4.2093
	COMP	0.2906	1.1465	1.3986	0.1757	1.1551	1.3648	0.1511	1.0083	1.0393
	MP	0.1768	0.7129	0.5394	0.1418	0.7582	0.5948	0.3328	0.6512	0.5347
	MI	0.0976	0.6594	0.4443	0.1002	0.6636	0.4504	0.4454	0.5591	0.5109
	MIMP	0.0971	0.6643	0.4506	0.1016	0.6580	0.4432	0.3223	0.4776	0.3319

DOI: 10.1002/sim

that adjusting for selection bias is very important. Since NAIVE is clearly not an option for analyzing such data, we will focus our discussion on the remaining four methods. The COMP method eliminates most of the bias in all simulation settings. This finding may appear to be in conflict with the known fact that ignoring missing values generally produces biased results when the response variables are missing. However, notice that in the propensity score estimation, it is covariates but not the outcome variables that have missing values. Removing observations with missing covariates is essentially the same as if only a subset of the data were collected. With the subset of the data, the regression coefficients are still unbiased as long as the data-generating model and the analysis model match. For setting A when data are MCAR, all four adjustment methods eliminate most of the bias. The MI and MIMP methods perform equally well with the smallest empirical biases, standard errors (estimated by standard deviations of simulated estimates), and mean-squared errors (MSEs). Although the COMP method is much less biased than the NAIVE estimator, it has much larger standard errors. The MP method has minimal biases, but it has larger standard errors than MI and MIMP. The MI and MIMP methods are approximately 10–30 per cent more efficient than MP where the relative efficiency is calculated as MSE of MP divided by MSE of MI or MIMP. For setting B when data are MAR, all four adjustment methods had minimal biases. Again, COMP has the largest standard errors and MP had larger standard errors than MI and MIMP. For $r=0.3, 0.5$ and 0.8, MI and MIMP have the smallest biases, standard errors and MSEs among all four adjusted methods and are approximately 30–50 per cent more efficient than MP. For $r=0$, MI and MIMP are slightly more biased than MP but they have smaller MSEs than MP (approximately 18 per cent more efficient). For setting C when data are MNAR, MI has larger biases than COMP, MP and MIMP. The standardized biases in the MI estimator were 0.07, 0.40, 0.72 and 1.53 for $r=0, 0.3, 0.5$ and 0.8, respectively, where the standardized bias is calculated as the empirical bias divided by the standard deviation of the NAIVE estimate. The MIMP method considerably reduced the bias compared with MI. The standardized biases in the MIMP estimator are 0.03, 0.27, 0.54 and 1.10 for $r=0, 0.3, 0.5$ and 0.8, respectively. As a result, MIMP has smaller MSEs and is 11, 26 and 54 per cent more efficient than MI for $r=0.3, 0.5$ and 0.8, respectively. The MP method performs well for setting C with similar bias as MIMP, but it has a larger standard deviation. As a result, MIMP is approximately 21–78 per cent more efficient than MP. The reason for the good performance of MP and MIMP is because the missingness pattern provides additional information for the missing values in the case of MNAR. In summary, for all simulation scenarios, MIMP and MI have the smallest MSE for setting A and B, and MIMP has the smallest MSE for setting C.

Table IV shows the simulation results from 5000 simulated samples for all settings with sample size $n=4000$. It is obvious that all standard deviations and MSEs are smaller than those for $n=500$. Again, in settings 1 and 2, MI and MIMP perform equally well with minimal biases and the smallest MSEs among all estimators for all r's. In setting 3, COMP had the smallest biases but large standard errors among all estimators. The biases in MP, MI and MIMP increase as r increases. MP and MIMP perform similarly and better than MI. For example, MIMP is 44, 69 and 218 per cent more efficient than MI for $r=0.3$, 0.5 and 0.8, respectively. The advantage of MIMP over MI is more dramatic for $n=4000$ compared with $n=500$. This is because for larger sample size, the biases have larger contribution to the MSEs as the standard errors of the estimators are smaller.

The MI and MIMP estimators in the simulation are based on $M=5$ imputations. From (10), it is easy to see that the superfluous variance due to only M imputations is $M^{-1}V_b$, and the proportion of such variance over the total variance is $M^{-1}V_b \div V_t$. In the simulation, V_t can be estimated

DOI: 10.1002/sim

Table IV. Empirical bias, standard deviation (SD) and mean-squared error (MSE) from 5000 simulations ($n=4000$).

		Setting A			Setting B			Setting C		
r	Method	Bias	SD	MSE	Bias	SD	MSE	Bias	SD	MSE
0	NAIVE	0.3229	0.0967	0.1136	0.3249	0.0971	0.1150	0.3207	0.0960	0.1121
	COMP	0.0045	0.1995	0.0398	0.0009	0.2117	0.0448	−0.0048	0.1957	0.0383
	MP	0.0001	0.1038	0.0108	0.0020	0.1075	0.0116	0.0012	0.1042	0.0108
	MI	0.0003	0.1022	0.0104	0.0019	0.1056	0.0111	0.0126	0.1023	0.0106
	MIMP	0.0003	0.1022	0.0104	0.0019	0.1056	0.0112	0.0007	0.1024	0.0105
0.3	NAIVE	1.1140	0.0993	1.2509	1.1124	0.1008	1.2477	1.1139	0.1002	1.2509
	COMP	0.0023	0.2841	0.0807	0.0065	0.2750	0.0757	−0.0012	0.2480	0.0615
	MP	0.0063	0.1455	0.0212	0.0055	0.1450	0.0211	0.0545	0.1370	0.0217
	MI	0.0053	0.1408	0.0198	0.0053	0.1386	0.0192	0.0993	0.1373	0.0287
	MIMP	0.0053	0.1409	0.0199	0.0054	0.1386	0.0192	0.0543	0.1304	0.0200
0.5	NAIVE	1.5156	0.1022	2.3075	1.5178	0.1015	2.3139	1.5195	0.0994	2.3189
	COMP	0.0206	0.3462	0.1203	0.0048	0.3495	0.1222	0.0109	0.3005	0.0904
	MP	0.0099	0.1910	0.0366	0.0075	0.1900	0.0362	0.1150	0.1591	0.0385
	MI	0.0101	0.1813	0.0330	0.0088	0.1819	0.0331	0.1776	0.1707	0.0607
	MIMP	0.0100	0.1815	0.0330	0.0085	0.1822	0.0333	0.1167	0.1490	0.0358
0.8	NAIVE	2.0254	0.1013	4.1126	2.0246	0.1037	4.1098	2.0263	0.1029	4.1165
	COMP	0.0422	0.5121	0.2639	0.0319	0.4548	0.2078	0.0255	0.4177	0.1751
	MP	0.0295	0.2829	0.0809	0.0176	0.2892	0.0839	0.2394	0.1956	0.0955
	MI	0.0177	0.2830	0.0804	0.0187	0.2739	0.0754	0.3987	0.2263	0.2102
	MIMP	0.0176	0.2834	0.0806	0.0188	0.2734	0.0751	0.2569	0.1744	0.0964

by the sample variance of the 5000 estimates, and V_b can be estimated by the mean of the $\hat{V}_b$ defined as (11) over 5000 simulations. For the MIMP estimator with $n=500$, $r=0.3$ and $M=5$ in setting B, the estimates for V_b and V_t are 0.0214 and $0.3997^2=0.1598$ (where 0.3997 is from Table III). Therefore, the superfluous variance due to only 5 imputations is $\frac{0.0214}{5}=0.00428$ and the proportion of such variance due to the limited number of imputations over the total variance is 2.7 per cent. In other words, the variance with infinite numbers of imputations could only be 2.7 per cent smaller than that with 5 imputations. This is consistent with Rubin's observation [9] that 3–5 imputations are generally sufficient to obtain excellent results.

4. APPLICATION TO CLINICAL DATA

In this section, we apply the five methods used for simulation to a set of data from an osteoporosis study: Multiple Outcome of Raloxifene Evaluation (MORE) [21]. In this study, 7705 women with osteoporosis were randomly assigned to one of the three treatment groups: placebo, raloxifene 60 mg/day or raloxifene 120 mg/day and were followed up for 4 years. After 3 years of follow-up, women were allowed to take other bone-active agents such as bisphosphonates. In this analysis, we compared the change in the femoral neck bone mineral density (BMD) during the fourth year (Y) between women not taking bisphosphonates (referred to as *untreated* group, i.e. $T=0$) and women taking bisphosphonates (referred to as *treated* group, i.e. $T=1$) for women who were originally

DOI: 10.1002/sim

Table V. The mean (95 per cent CI) of the change in femoral neck BMD (g/cm^2) during the fourth year of MORE study for women who were assigned to placebo ($n=1643$).

Method	Untreated ($n=1512$)	Treated ($n=131$)	Treated vs untreated
NAIVE	$-0.002(-0.004, -0.001)$	$0.011(0.006, 0.017)$	$0.013(0.007, 0.019)$
COMP	$-0.002(-0.004, -0.001)$	$0.006(-0.002, 0.020)$	$0.008(0.000, 0.022)$
MP	$-0.002(-0.004, -0.001)$	$0.007(-0.003, 0.016)$	$0.009(-0.001, 0.018)$
MI	$-0.002(-0.003, -0.001)$	$0.009(0.002, 0.017)$	$0.011(0.004, 0.019)$
MIMP	$-0.002(-0.003, -0.001)$	$0.009(0.002, 0.017)$	$0.011(0.004, 0.019)$

assigned to placebo. Because taking bisphosphonates was not a randomized factor, the response variable Y in treated and untreated groups may be confounded by the selection bias. We estimated the treatment effect using the propensity score approach including a total of 16 covariates: age at baseline (i.e. prior to randomization), body mass index at baseline, family history of breast cancer, 5-year breast cancer risk score [22] at baseline, whether a women was hysterectomized at baseline, lumbar spine BMD at baseline, femoral neck BMD at baseline, change in lumbar spine BMD during the first 3 years, change in femoral neck BMD during the first 3 years, nonvertebral fracture prior to baseline, prevalent vertebral fracture at baseline, new nonvertebral fracture in the first 3 years, new vertebral fracture in the first 3 years, weighted adverse event score during the first 3 years calculated as ($1\times$#mild AE$+2\times$#moderate AE$+3\times$#severe AE$+4\times$#serious AE), smoking status at baseline, baseline semi-quantitative vertebral fracture status ($0=$no fracture, $1=$ mild fracture, $2=$moderate fracture and $3=$severe fracture).

Data were analyzed using SAS 9.1.3. The Markov chain Monte Carlo method [23] in SAS procedure 'PROC MI' was used to perform multiple imputation. There were a total of 1643 ($n=1512$ for $T=0$ and $n=131$ for $T=1$) women included in this analysis and 603 (36.7 per cent) women had at least one missing covariate. There were 14 missingness patterns. The largest pattern had 1040 subjects and each of the 3 smallest patterns had only 1 observation. We pooled pattern with $d_{\min}=100$. There were 2 combined patterns with 1087 and 556 observations. Table V shows the estimation results based on the five methods: NAIVE, COMP, MP, MI and MIMP. The 95 per cent confidence interval (CI) was estimated by the BCa bootstrap method with 1000 resamplings. The variance used in the acceleration parameter was estimated by the jacknife method. The difference of the femoral neck BMD changes between treated and untreated groups by the NAIVE method appeared higher than the estimates by other methods. The MI and MIMP methods produced similar results and had much narrower CIs than the COMP method, but the point estimates from MI and MIMP were close to the NAIVE estimator. The MP estimate and the 95 per cent CI were similar to those from COMP. This is consistent with the simulation results in that MI and MIMP had smaller standard errors and MSEs, COMP had the smallest bias, and MP had smaller bias than MI and MIP.

5. SUMMARY AND DISCUSSION

The propensity score approach is used widely to reduce bias in estimating the treatment effect for nonrandomized studies. As there can be a large amount of missing data, proper handling of

DOI: 10.1002/sim

missing values in estimating propensity scores is important. In this article, we proposed to use multiple imputation missingness pattern (MIMP) to estimate the propensity scores. The MIMP method is easy to implement in standard statistical software packages such as SAS, S-PLUS and R. Simulation shows that MIMP and MI perform approximately equally well and produce almost unbiased estimators for MCAR and MAR. For MNAR, the MIMP method modestly outperforms the MI method, especially when the correlations among the covariates are large. Although we used the IPW estimator in utilizing propensity scores in our simulations and the data example, clearly the MIMP method should also work for other propensity score-based methods.

Several practical issues should be considered when applying the MIMP method to a real data problem. The standard error of the point estimator based on MI and MIMP could not rely on the simple Rubin's formula combining within and between imputation variance, because the standard error estimated from the general IPW estimator does not account for the uncertainty of the estimated weights. Therefore, resampling techniques such as bootstrap or jackknife method may be used for inference purpose. In contrast to Tu and Zhou's conclusion [17], we found that the general BCa bootstrap method can be applied directly to the propensity score-based approaches using MI and MIMP.

Although simulation shows that MIMP performs similar to MI for MCAR and MAR, MIMP is a more robust method for MNAR. MIMP should always be considered because in practice, there is no way to verify the missingness mechanism and in fact MNAR may often occur for real data. For example, in clinical trials or observational studies, patients who have poorer efficacy tend to be noncompliant (e.g. missing clinic visits). In this situation, missingness mechanism is not at random because whether patients miss the evaluation visit depends on the efficacy measurements but the efficacy measurements could be missing due to missing clinic visits.

MIMP utilizes the missingness pattern indicator and may produce unstable estimates when some patterns have small number of observations. To address this problem, we proposed an algorithm to pool small patterns to form new cells which have the minimum number of observations in each cell and roughly have the same number of observations across pooled cells. In practice, the minimum threshold value for the number of observations in each pooled cells needs to be determined according to the number of covariates and the distribution of treatment assignment. For example, in the simulation with 12 covariates and approximately 50 per cent of subjects on each treatment, pooling patterns with $d_{\min}=50$ generally provide good stable estimates. In the application with 16 variables and less than 10 per cent of subjects untreated (Section 4), for $d_{\min}=50$, the propensity score model failed to converge. Therefore, $d_{\min}=100$ was used.

There are other ways to utilize the missingness pattern information. For example, in addition to the variables in MIMP, similar to MP, one could include the terms for the interactions between the indicator variables for the patterns of missing data and the covariates in the propensity score estimation model after multiple imputation. Although including interaction terms in the model uses more information about the missingness pattern, it increases the number of parameters to estimate and generally results in inefficiency in the estimation of propensity scores.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Douglas Faries and Dr Brenda Crowe for their many helpful comments and suggestions for this research. We would also like to thank the two anonymous referees for their useful comments and Dr Michelle S. Lewis for her careful editorial reviews. These comments and suggestions have led to significant improvement of this manuscript.

DOI: 10.1002/sim

REFERENCES

1. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *The New England Journal of Medicine* 2000; **342**:1887–1892.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
3. D'Agostino R. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998; **17**:2265–2281.
4. D'Agostino R, Lang W, Walkup M, Morgon T. Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services and Outcomes Research Methodology* 2001; **2**:291–315.
5. Ibrahim J, Lipitz S, Chen M. Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *Journal of the Royal Statistical Society*, *Series B* (*Statistical Methodology*) 1999; **61**:173–190.
6. D'Agostino R, Rubin DB. Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 2000; **95**:749–759.
7. D'Agostino R. Propensity score estimation with missing data. In *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, Gelman A, Meng X-L (eds). Wiley: New York, 2001; 163–174.
8. Rubin DB. Multiple imputations in sample surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association* 1978; 20–34.
9. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, 1987.
10. Crowe B, Lipkovich I, Wang O. Comparison of several imputation methods for missing baseline data in propensity scores analysis of binary outcome. *Pharmaceutical Statistics*; under review.
11. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Medical Association* 1952; **47**:663–685.
12. Rosenbaum PR. Model-based direct adjustment. *Journal of the American Statistical Association* 1987; **82**:387–394.
13. Xie J, Liu C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine* 2005; **24**:3089–3110.
14. Lunceford JK, Davidian M. Stratification sand weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* 2004; **23**:2937–2960.
15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846–866.
16. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Wiley: New York, 2002.
17. Tu W, Zhou X-H. Bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *UW Biostatistics Working Paper Series*, University of Washington, The Berkeley Electronic Press (bepress). Available from: http://www.bepress.com/uwbiostat/paper200.
18. Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 1987; **82**:171–200.
19. Van Buuren S, Oudshoorn CGM. Flexible multivariate imputation by MICE. *Report PG/VGZ/99.054*, TNO Prevention and Health, Leiden, 1999.
20. Van Buuren S, Oudshoorn CGM. Multivariate imputation by chained equations: MICE V1.0 user's manual. *Report PG/VGZ/00.038*, TNO Prevention and Health, Leiden, 2000.
21. Delmas PD, Ensrud KE, Adachi JD, Harper KD, Sarkar S, Gennari C, Reginster J-Y, Pols HAP, Recker RR, Harris ST, Wu W, Genant HK, Black DM and Richard Eastell for the Multiple Outcomes of Raloxifene Evaluation (MORE) Investigators. Efficacy of raloxifene on vertebral fracture risk reduction in postmenopausal women with osteoporosis: four-year results from a randomized clinical trial. *The Journal of Clinical Endocrinology and Metabolism* 2002; **87**(8):3609–3617.
22. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J, Wieand HS. Validation studies for models projecting the risk of invasive and total breast cancer incidence. *Journal of the National Cancer Institute* 1999; **91**(18):1541–1548.
23. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: New York, 1997.

DOI: 10.1002/sim