

MANUSCRIPT

A multiple imputation-based sensitivity analysis approach for data subject to missing not at random

Chiu-Hsieh Hsu¹ | Yulei He² | Chengcheng Hu¹ | Wei Zhou³

¹Department of Epidemiology and Biostatistics, College of Public Health, University of Arizona, Tucson, AZ, USA

²National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA

³Department of Surgery, University of Arizona, Tucson, MI, USA

Correspondence

Chiu-Hsieh Hsu, Department of Epidemiology and Biostatistics, University of Arizona, 1295 N Martin Ave, Tucson, AZ 85724.

Email: pchhsu@email.arizona.edu

Abstract

Missingness mechanism is in theory unverifiable based only on observed data. If there is a suspicion of missing not at random (MNAR), researchers often perform a sensitivity analysis to evaluate the impact of various missingness mechanisms. In general, sensitivity analysis approaches require a full specification of the relationship between missing values and missingness probabilities. Such relationship can be specified based on a selection model, a pattern-mixture model or a shared parameter model. Under the selection modeling framework, we propose a sensitivity analysis approach using a nonparametric multiple imputation strategy. The proposed approach only requires specifying the correlation coefficient between missing values and selection (response) probabilities under a selection model. The correlation coefficient is a standardized measure and can be used as a natural sensitivity analysis parameter. The sensitivity analysis involves multiple imputations of missing values, yet the sensitivity parameter is only used to select imputing/donor sets. Hence, the proposed approach might be more robust against misspecifications of the sensitivity parameter. For illustration, the proposed approach is applied to incomplete measurements of level of pre-operative Hemoglobin A1c, for patients who had high-grade carotid artery stenosis and were scheduled for surgery. A simulation study is conducted to evaluate the performance of the proposed approach.

KEYWORDS:

Correlation coefficient, Missing not at random, Multiple imputation, Selection model, Sensitivity analysis

1 | INTRODUCTION

Missing data problems are common in biomedical studies, and thus could lead to a substantial bias and misleading inference if handled inappropriately. Most existing statistical methods analyze incomplete data under the missing completely at random (MCAR) or missing at random (MAR) assumption. In some situations, the missing not at random (MNAR) assumption might

be more plausible. For example, in electronic medical record databases, patients with normal Hemoglobin A1c (HbA1c) level are less likely to have HbA1c tests performed, and therefore have missing HbA1c records in the database. This fact might imply a MNAR mechanism in some analyses: using data only from those who have their HbA1c level recorded is likely to over-estimate the average HbA1c level of the general patient population. However, the underlying missingness mechanism is in theory unverifiable using the information only from observed data. Hence, it is important to perform sensitivity analyses that assess the impact on study results from different missingness mechanism assumptions.

Multiple sensitivity analysis approaches have been proposed in the literature. A basic strategy is the delta adjustment procedure¹, from which the extensions include pattern mixture² and tipping point³ sensitivity analysis approaches. The delta adjustment is directly applied to the value of the missing variable or the missingness probability. For example, suppose that we are interested in estimating the mean of a single incomplete variable Y without any covariate. When the adjustment is applied to the value of the missing Y , a simple sensitivity analysis parameter δ , which determines the potential missingness mechanism to be MCAR (i.e. $\delta = 0$) or MNAR (i.e. $\delta \neq 0$), can be used to quantify the average difference between missing and observed cases. Suppose that the average of observed values for this variable is μ_{obs} , then the average of the full data set might be expressed as $\mu = p_{obs}\mu_{obs} + (1 - p_{obs})(\mu_{obs} + \delta) = \mu_{obs} + (1 - p_{obs})\delta$, where p_{obs} is the proportion of respondents. In this simple case, suppose that a larger $|\delta|$ implies a stronger MNAR mechanism (note that MCAR would imply $\delta = 0$). Specifying a range of δ 's and then obtaining the corresponding μ 's would provide some ideas about the sensitivity of the full-data inference under various MNAR mechanisms. This idea can be extended to include covariates¹. On the other hand, the adjustment can be applied to the missingness probability, in which the sensitivity parameter quantifies the relationship between the missing variable and the missingness probability.

However, based on our knowledge, sensitivity parameters are usually not standardized. Therefore, it is not straightforward to characterize the magnitude of MNAR in these sensitivity analyses. For instance, in the aforementioned example, the difference between the means of missing and observed cases, δ , can theoretically range from $-\infty$ and ∞ , which might not be very informative. Heckman's selection model^{4,5} uses the correlation coefficient between the variable subject to missingness and an unobservable latent variable associated with the missingness probability to specify the missingness mechanism. The correlation coefficient is always between -1 and $+1$ and standardized. Heckman's selection model has been used to develop a sensitivity analysis approach through a profile likelihood method⁶, in which the sensitivity analysis parameter is a function of the correlation coefficient and standardized. However, to perform this sensitivity analysis approach, it requires estimating the bivariate normal distribution between the variable subject to missingness and the unobservable latent variable even if the correlation coefficient is pre-specified. In other words, the sensitivity parameter is directly incorporated into the estimation. Also, the estimation of Heckman's selection model is very sensitive to the normality assumption, which is often violated in practice, and high correlation between the missing variable and the latent variable. Therefore, one would suspect the sensitivity analysis approach

based on Heckman's selection model will be highly sensitive to mis-specification of the model. Hence, it is of interest to develop a sensitivity analysis approach that has a standardized sensitivity parameter indicating the magnitude and direction of MNAR and the sensitivity parameter is not directly incorporated into estimating the value of the missing variable or the missingness probability under a selection model framework.

For missing data under a MAR assumption, a nonparametric multiple imputation method using information from additional fully observed covariates based on a nearest-neighbor approach was previously proposed^{7,8}. This approach develops a distance measure between the incomplete and observed cases, using a weighted sum of predictive scores derived from the two working models: one for predicting the missing outcome and the other for predicting the missingness probability. An imputing set/donor pool (i.e. observed cases with the closest distance) is selected for each missing observation, and all of the observations in the imputing set have an equal chance to be drawn to replace the missing observation. This strategy can be viewed as an extension of the predictive mean-matching/hot deck multiple imputation^{9,10}, in which the distance metric is derived only from the outcome working model. An advantage of this approach is that, since the working models are only used to identify the imputing set, it might be more robust against misspecifications of working models than the imputation methods which directly use working models for generating imputations.

In this paper, we will modify the nonparametric multiple imputation method to develop a sensitivity analysis approach under a selection modeling framework. The sensitivity parameter is the correlation coefficient (ρ) between the missing variable and the selection/response probability in Heckman's selection model, which is standardized. Similar to the delta adjustment procedure, the sensitivity parameter ρ determines the potential missingness mechanism to be MCAR (i.e. $\rho = 0$) or MNAR (i.e. $\rho \neq 0$). Unlike the delta adjustment procedure or any parametric sensitivity analysis approach, the sensitivity parameter ρ in our approach is not directly applied to the value of the missing variable or the missingness probability or incorporated into estimation of the parameters but only used to identify imputing sets for missing observations. In this regard, our approach relies less on the selection model. The remainder of this paper is organized as follows. Section 2 first describes the setup and Heckman's selection model and then describes the sensitivity analysis procedure. Section 3 assesses the performance of the approach through simulation studies. Section 4 presents a real application. Finally, Section 5 concludes with a discussion and points out directions for future studies.

2 | METHOD

2.1 | Selection Model

This section presents a brief introduction of Heckman's selection model^{4,5}. For simplicity, we first consider the situation with no additional fully observed covariates. Let Y denote the variable subject to missingness and S denote the selection/response

indicator, i.e. $S = 1/0$ if Y is observed/missing. Under the selection modeling framework, the joint density function of Y and S can be expressed as

$$f_{Y,S}(Y, S) = f_Y(Y; \theta)P_{S|Y}(S|Y; \psi) \quad (2.1)$$

where $f_Y(\cdot)$ is the density function of Y and θ is the associated parameter, and $P_{S|Y}(\cdot)$ is the conditional probability function of S on Y and ψ is the associated parameter. Note that under MAR (or ignorable missingness), $P_{S|Y}(S|Y; \psi)$ can be ignored in the estimation process so that the inference for θ can be made only based on $f_Y(Y; \theta)$.

Under Eq. (2.1), MNAR can be induced through an unobservable latent variable X to correlate the selection probability with Y . Specifically,

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim BVN(\mu, \Sigma) \\ Pr(S = 1) = Pr(X > 0) \quad (2.2)$$

where BVN stands for a bivariate normal distribution, $\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}$, $\Sigma = \begin{bmatrix} \sigma_y^2 & \rho\sigma_y \\ \rho\sigma_y & 1 \end{bmatrix}$, and ρ is the correlation coefficient between Y and X . Model (2.2) is known as the Heckman's selection model^{4,5}. The correlation parameter ρ controls the magnitude of MNAR: $\rho = 0$ indicates that missingness is MCAR; $\rho \neq 0$ indicates that missingness is MNAR. A positive (negative) ρ suggests that cases with smaller (larger) values are more likely to be missing.

Under Model (2.2), it can be shown that

$$f(Y|S = 1) = \frac{1}{\Phi(\mu_x)} \phi\left(\frac{Y - \mu_y}{\sigma_y}\right) \Phi\left(\frac{\mu_x + \rho\left(\frac{Y - \mu_y}{\sigma_y}\right)}{\sqrt{1 - \rho^2}}\right) \quad (2.3)$$

$$f(Y|S = 0) = \frac{1}{\Phi(-\mu_x)} \phi\left(\frac{Y - \mu_y}{\sigma_y}\right) \Phi\left(\frac{-\mu_x - \rho\left(\frac{Y - \mu_y}{\sigma_y}\right)}{\sqrt{1 - \rho^2}}\right), \quad (2.4)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ stand for the probability density and cumulative distribution functions for the standard normal distribution, respectively. Assuming that all the parameters of Model (2.2) are known, Eq. (2.3) and (2.4) state the probability density distribution of observed cases Y_{obs} and missing cases Y_{mis} , respectively.

In addition, the conditional mean of Y on $S = 1$ and $S = 0$, respectively, can be expressed as $E[Y|S = 1] = \mu_y + \frac{\phi(\mu_x)}{\Phi(\mu_x)}\rho\sigma_y$ and $E[Y|S = 0] = \mu_y + \frac{-\phi(\mu_x)}{1 - \Phi(\mu_x)}\rho\sigma_y$. This indicates that the bias of the complete case (CC) analysis depends on the inverse Mills ratio (i.e. $\frac{\phi(\mu_x)}{\Phi(\mu_x)}$) and increases with $|\rho|$ and σ_y . When ρ is positive, CC tends to over-estimate $E(Y)$, the population mean. When ρ is negative, CC tends to under-estimate $E(Y)$.

In theory Model (2.2) cannot be estimated without the inclusion of additional fully observed covariates for both Y and S . In practice, often there are additional covariates observed, say Z_1, \dots, Z_p , which might be correlated with the missing outcome or the missing probability and can then be incorporated into $\mu_y (= \beta_0 + \beta_1 Z_1 + \dots + \beta_p Z_p)$ and $\mu_x (= \alpha_0 + \alpha_1 Z_1 + \dots + \alpha_p Z_p)$ in Model (2.2). Therefore, selection models with fully observed covariates are often used for the purpose of estimation, not for performing sensitivity analysis. On the other hand, the sensitivity analysis strategy proposed in this paper focuses on specifying (not estimating) ρ in handling missing data subject to MNAR. Hence the related discussion still centers on Model (2.2). Extensions to scenarios with covariates can be found in Step 3 of the imputation procedures in Section 2.3, where ρ is the correlation coefficients between $\epsilon_y = Y - \mu_y$ and $\epsilon_x = X - \mu_x$ and also indicates the missing mechanism. Specifically, when $\rho = 0$, it is MCAR if Y and X are independent and it is MAR if Y and X are conditionally independent. When $\rho \neq 0$, it is MNAR even after conditional on Z . To correct the bias associated with CC using the delta adjustment procedure where $\delta = -\frac{\phi(\mu_x)}{\Phi(\mu_x)}\rho\sigma_y$, one needs to specify the inverse Mills ratio, ρ and σ_y . In contrast, the proposed approach only needs to specify ρ to select imputing sets based on the order of the observed data (i.e. $S = 1$) to correct bias associated with CC.

2.2 | Multiple Imputation

Multiple imputation¹¹ is arguably the most popular statistical strategy to practical missing data analyses. For illustration, suppose that $Y = (Y_{mis}, Y_{obs})$, where Y_{mis} and Y_{obs} are the missing and observed part of Y , respectively. Assuming MAR for Y_{mis} , then a Bayesian, model-based multiple imputation analysis consists of two main steps. The first step draws imputations for Y_{mis} from its posterior predictive distribution, $f(Y_{mis}|Y_{obs})$, independently multiple (say M) times to create M sets of completed data. Drawing imputations typically involves the estimation of θ in Eq. (2.1). In the second step, complete-data analysis procedures (e.g., means and regressions) are applied to these M completed datasets to create M sets of analysis results, and these results are combined to yield a single set of inference (e.g., point and variance estimates, confidence intervals and p -values, etc) using so-called Rubin's combining rules. There is a large body of literature about the theory and applications of multiple imputation. Some classic literature, for example, can be found in¹¹ and¹².

2.3 | Multiple-Imputation Based Sensitivity Analysis

For missing data under MNAR, multiple imputations for missing data need to be drawn from $f(Y_{mis}|Y_{obs}, S)$, where S is the response/sampling indicator in Eq. (2.1). Multiple imputation has been used to develop sensitivity analysis approaches based on both pattern-mixture¹³ and selection models. Specifically, parametric multiple imputation approaches based on the Heckman's selection model and its extensions have been proposed^{14,15}. Another strategy uses a selection model to first induce MNAR and then develop dual imputation procedures to iteratively impute both missing indicator and missing variable¹⁶. These approaches involve a direct estimation of parameters of the selection model. Yet estimation of selection models can be rather sensitive to the

model specification¹¹. In contrast, we propose a sensitivity analysis strategy. Specifically, it specifies a range of ρ , the correlation coefficient between Y and X , and use ρ and Y_{obs} to generate imputations for missing observations. We then obtain the multiple imputation analysis results for each ρ , and hence can gain insights about how sensitive the results are to the extent of MNAR.

We now discuss how the imputing set is determined for specified ρ and Y_{obs} . Eqs (2.3) and (2.4) imply that

$$f(Y_{mis}) = f(Y_{obs}) \frac{\Phi(\mu_x)}{1 - \Phi(\mu_x)} \frac{1 - \Phi\left(\frac{\mu_x + \rho\left(\frac{Y - \mu_Y}{\sigma_Y}\right)}{\sqrt{1 - \rho^2}}\right)}{\Phi\left(\frac{\mu_x + \rho\left(\frac{Y - \mu_Y}{\sigma_Y}\right)}{\sqrt{1 - \rho^2}}\right)}, \quad (2.5)$$

where $f(Y_{mis}) = f(Y|S = 0)$ and $f(Y_{obs}) = f(Y|S = 1)$. Note that $\frac{\Phi(\mu_x)}{1 - \Phi(\mu_x)}$ is a constant for a given μ_x . Therefore conditioning

on all the parameters, the posterior distribution of Y_{mis} can be viewed as that of Y_{obs} weighted by the factor, $W = \frac{1 - \Phi\left(\frac{\mu_x + \rho\left(\frac{Y - \mu_Y}{\sigma_Y}\right)}{\sqrt{1 - \rho^2}}\right)}{\Phi\left(\frac{\mu_x + \rho\left(\frac{Y - \mu_Y}{\sigma_Y}\right)}{\sqrt{1 - \rho^2}}\right)}$.

Apparently, for a positive ρ , $W \rightarrow 0$ as $Y \rightarrow \infty$ and W increases as $Y \rightarrow -\infty$. The trend of W is reversed for a negative ρ . This suggests that to find an appropriate replacement of a missing observation, larger (smaller) observed values should have a lower (higher) probability (i.e., weight) to be used for a positive (negative) ρ .

A direct imputation approach would draw missing values from Eq. (2.5). In the proposed sensitivity analysis, on the other hand, we define a neighbourhood of Y_{obs} as the imputing set. For a positive ρ , the neighbourhood would be the bottom $100(1 - \rho)\%$ of the Y_{obs} -values, downweighting relatively larger Y_{obs} -values. For a negative ρ , the neighbourhood would be the top $100(1 + \rho)\%$ of the Y_{obs} -values, downweighting relatively smaller Y_{obs} -values. When $\rho = 0$, which is MCAR, the imputing set includes all Y_{obs} -values. The detailed imputation algorithm is described as follows:

Step 1: Generate a bootstrap sample of the original dataset. A sampling with replacement will be performed to generate a bootstrap sample with the same size as the original dataset. This step incorporates the uncertainty of parameter estimates and will result in proper multiple imputation¹⁷.

Step 2: Specify a correlation coefficient, ρ , between Y and X . For each specified ρ , a ρ^* will be drawn from the asymptotic distribution, $N\left(\frac{1}{2} \ln \frac{1 + \rho}{1 - \rho}, (n - 3)^{-1}\right)$, based on the Fisher transformation. The sampling step is to incorporate the uncertainty of ρ in the multiple imputation process. The sampled ρ^* will be used in the later steps. Steps 3 and 4 will be performed for the range of ρ^* 's. However, for the easiness of the notation, we do not distinguish between ρ and ρ^* hereafter.

Step 3: Define the imputing set. For each subject j with a missing outcome Y in the original dataset, ρ in Step 2 can be employed to define the size of nearest neighbors (NN) (i.e. imputing set). Under a situation with no additional fully observed covariates available, NN is defined as $n_{obs} \times (1 - |\rho|)$, where n_{obs} is the number of subjects with Y observed in the bootstrap sample. This indicates that the neighborhood, $R(NN)$, consists of $100 \times (1 - |\rho|)\%$ of n_{obs} subjects who have their Y observed in the bootstrap sample. Under a situation with additional fully observed covariates available, NN is defined as $n_{obs} \times (1 - |\rho|)$, where n_{obs} is the number of subjects in the bootstrap sample with Y observed and similar covariate values as subject j . Specifically,

for a categorical covariate Z , the subjects included in NN needs to be in the same category as subject j . For a continuous covariate Z , the covariate will be first categorized into groups based on its percentiles and then used to select subjects in the bootstrap sample classified into the same percentile group as subject j to prevent from having empty imputing sets for some missing subjects. This indicates that the neighborhood, $R(NN)$, consists of $100 \times (1 - |\rho|)\%$ of n_{obs} subjects in the bootstrap sample who have their Y observed and similar covariate values as subject j . Based on the way that NN is constructed, when $\rho > 0$ (i.e. a case of MNAR with smaller values of Y more likely to be missing, the neighborhood $R(NN)$ includes the smallest observed Y up to the top $100 \times (\rho)^{th}$ percentile of the observed Y . When $\rho < 0$ (i.e. a case of MNAR with larger values of Y more likely to be missing, the neighborhood $R(NN)$ includes the top $100 \times (1 + \rho)^{th}$ percentile of the observed Y up to the largest observed Y . For example, when $\rho = 0.2$, $R(NN)$ consists of the (bootstrap) Y_{obs} ranging from the minimum to its 20th percentile. When $\rho = -0.2$, $R(NN)$ consists of the (bootstrap) Y_{obs} ranging from its 80th percentile to the maximum. When $\rho = 0$ (i.e. a case of MCAR), $R(NN)$ consists of all of the (bootstrap) Y_{obs} when there is no additional fully observed covariates and consists of the entire NN when there are additional fully observed covariates. When $\rho = 1$, the minimum in NN will be imputed for each missing observation. When $\rho = -1$, the maximum in NN will be imputed for each missing observation.

Step 4: Impute a value from the imputing set. For any missing subject, a value is randomly drawn from the imputing set, $R(NN)$, to replace the missing value. Each value in $R(NN)$ has an equal probability to be drawn.

Step 5: Repeat Steps 1 to 4 independently M times and combine the multiple imputation analysis results. The procedure can be independently repeated M times to obtain multiple imputed data sets for use in estimation. Once the M multiply imputed datasets are obtained, we carry out the multiple imputation analysis procedure established in¹ for a specified ρ . For example, the marginal mean of Y (i.e. μ_y) will be estimated on the M imputed datasets. The final estimate (i.e. $\hat{\mu}_y$) is the average of the M sample mean (i.e. \bar{Y}) and the final variance is the sum of a between-imputation and a within-imputation component¹¹. Similarly, the regression coefficients (i.e. β) will be estimated by fitting a linear regression model to each of the M imputed datasets. The final estimate for a specific regression coefficient, e.g. β_k , is the average of the M $\hat{\beta}_k$ and the final variance is the sum of a between-imputation and a within-imputation component¹¹.

Step 6: Repeat Steps 1 to 5 for a range of specified ρ 's and display the corresponding multiple imputation analysis results obtained in Step 5. Since the goal is to conduct a sensitivity analysis, there is no strong rationale for choosing specific ρ 's. We can vary ρ from very negative (i.e., close to -1) to very positive (i.e., close to 1) values, as well as including 0.

3 | SIMULATION

We perform several simulation studies to investigate the finite-sample performance of the proposed sensitivity analysis strategy. In the simulation, we investigate effects of several factors on the performance: the magnitude of MNAR (measured by ρ), the

variance of Y (i.e. σ_y^2), the range of specified ρ , and the true distribution of Y and X . The sample code for the simulation is written in R and is available upon request.

The simulation consists of 500 replicates. In Scenario I, we assume that Model (2.2) holds and then generate Y and X from the selection model. In the data generation process, we set $\mu_x = 0.5$ to yield a missingness rate approximately 31% for Y . We also consider $\sigma_y = (0.8, 1.0, 1.2)$ and choose ρ ranging from -0.50 to 0.50 . The sensitivity analysis specifies the ρ using the true ρ .

In Scenario II, the data are generated similar to that in Scenario I, except that the ρ is fixed at -0.30 and 0.30 . In the sensitivity analysis, we specify a range of ρ 's, which can deviate from the true ρ by ± 0.25 . More specifically, for true $\rho = -0.30(0.30)$, the specified ρ ranges from $-0.55(0.05)$ to $-0.05(0.55)$.

In Scenario III, we do not generate data from Model (2.2). Instead, we first generate data from two correlated standard normal variables. Once the two variables are generated, their percentiles under the standard normal distribution are derived and used to identify the values with the same percentiles under a pre-specified exponential distributions with hazard rates λ_y and λ_x , respectively. Hence, the Y and X in this scenario follow an exponential distribution with a hazard rate of λ_y and λ_x , respectively, and they are also correlated. The selection indicator is still defined as $S = I(X > c)$, where c is used to give a missingness rate around 30% for Y , i.e. $c = -\log(0.7)/\lambda_x$. We calculate the Pearson correlation coefficient r , Spearman correlation coefficient s , and Kendall's τ using Y and X . In practice, those correlation coefficients cannot be calculated because X is unobservable. In the sensitivity analysis, we test and compare the performance of three different specified ρ 's by setting them as these calculated statistics, respectively, to determine which correlation coefficient (if it is known) as a sensitivity analysis parameter has a better performance when the data are not generated from the selection model and are not normally distributed.

In Scenario IV, the data are generated from Model (2.2) with additional two fully observed covariates, i.e. Z_1 and Z_2 , where Z_1 is generated from *Bernoulli*(0.5) and Z_2 is generated from *Uniform*(0, 1), $\mu_y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$, $\mu_x = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2$, and $\sigma_y = 1.0$. We set $\alpha_0 = 0.3$, $\alpha_1 = 0.5$ and $\alpha_2 = -0.5$ to control the missingness rate approximately 39% and choose ρ ranging from -0.50 to 0.50 .

We apply the following missing data analysis methods to the simulated data:

1. The "Fully-Observed" (FO) analysis conducts the estimation before any missingness is applied. The FO analysis is the gold standard.
2. The complete-case (CC) analysis conducts the estimation based on the respondents (i.e. $S = 1$).
3. The predictive mean matching multiple imputation method (PMM-MI) conducts the estimation based on the predictive mean matching imputed datasets.

4. The proposed sensitivity analysis approach (NNMI) follows the procedure listed in Section 3. Multiple imputation analysis is based on $M = 5$ imputed datasets. In a situation with two additional fully observed covariates available, i.e. Z_1 and Z_2 , the continuous covariate Z_2 is categorized into g groups based on the percentiles (denoted as NNMI_g).

Tables 1-3 show the main simulation results. The FO analysis yields little biases in all situations and produces coverage rates around the nominal level, 95%. As expected, the CC analysis produces biased estimates and has a lower-than-nominal coverage rate in all situations. In addition, the bias for CC increases as $|\rho|$ and σ_y increase, consistent with the algebraic arguments at the end of Section 2.1.

Table 1 shows the results when ρ is correctly specified in the sensitivity analysis (Scenario I). As pointed out in Section 2.3, the NNMI method does not directly impute missing values based on the selection model. Therefore we do not expect that its performance would perfectly align with that of FO analysis. However, the biases of NNMI are overall small, and their coverage rates are also close or comparable to the nominal level, even for situations with a high degree of MNAR (i.e. $|\rho| = 0.50$). Similar to CC, the bias for NNMI also increases with $|\rho|$ and σ_y increase, yet the gradient of increase is much smaller than that of CC. For example, when $\sigma_y = 0.8$ and ρ increases from 0 to 0.3, the bias increases from 0.002 to 0.008 for NNMI and yet from 0.002 to 0.123 for CC.

Table 2 shows the results when the specified ρ deviates from the true ρ (Scenario II). As expected, as the specified ρ moves away from the true ρ , the corresponding results are more biased and with lower coverage rates. When the specified ρ does not deviate more than 0.15 from the true ρ , the NNMI results are satisfactory or acceptable, and they fare much better than CC. For example, when the specified ρ is -0.40, 33% lower than the true $\rho = -0.30$, the bias of NNMI increases from -0.003 to 0.032, 0.040 and 0.048 for $\sigma_y = 0.80, 1.0$ and 1.2 , respectively, which are much smaller than the corresponding biases associated with CC (i.e. -0.119, -0.148 and -0.177). When the specified ρ is 0.40, 33% higher than the true $\rho = 0.30$, the bias increases from 0.006 to -0.029, -0.036 and -0.043 for $\sigma_y = 0.80, 1.0$ and 1.2 , respectively, which are much smaller than the corresponding biases associated with CC (i.e. 0.123, 0.154 and 0.185). In addition, the coverage rates of NNMI are slightly off from the nominal level, when the specified ρ is 33% lower or higher than the true ρ , especially when the true $\rho < 0$. When the specified ρ deviates 0.25 from the true ρ , the bias of NNMI is still lower than bias associated with CC.

Table 3 shows the results when Y and X are not generated from the selection model (Scenario III). As expected, the results from NNMI fare slightly worse than those in the scenario where Y and X are indeed generated from a selection model (e.g., Table 4). Interestingly, here NNMI using Spearman correlation coefficient, ρ , (i.e. $\text{NNMI}_{\rho=s}$) has overall a better performance than using Pearson correlation coefficient or Kendall's τ (i.e. $\text{NNMI}_{\rho=r}$ and $\text{NNMI}_{\rho=\tau}$).

Table 4 shows the results of regression analysis when two additional fully observed covariates are available and incorporated into selecting NN (Scenario IV). As expected, the FO analysis yields little biases in all regression coefficient estimation and produce coverage rates around the nominal level, 95%. The biases of CC and PMM-MI increase as the magnitude of MNAR

increases, especially for β_0 . In general the biases of NNMI are smaller than CC, which relies on the MCAR assumption, and PMM-MI, which relies on the MAR assumption, especially for β_0 . The results also indicate that categorizing the continuous covariate Z_2 into 4 or 5 groups is sufficient to produce reasonable estimates for all three regression coefficients except when $\rho = 0$. When $\rho = 0$, i.e. MCAR, a larger number of group is needed to produce reasonable estimates, especially for β_0 .

In summary, the bias of CC analysis increases with ρ and σ_y when missingness is MNAR. On the other hand, NNMI using ρ to select imputing sets for missing observations in the proposed sensitivity analysis can correct some biases of CC analysis even if the specified ρ considerably (e.g. $\pm 10\%$ of the true ρ) deviates from the true ρ . In addition, the collection of results provides some insights how the missing data inference is affected by the magnitude of MNAR. Finally, even if the data do not follow the selection model, the sensitivity analysis strategy seems to yield acceptable results.

4 | APPLICATION

For illustration, we apply the sensitivity analysis approach to a clinical dataset measuring patients' HbA1c level. This illustrative dataset consists of patients who had high-grade carotid artery stenosis and were scheduled to undergo carotid artery interventions at the Veterans Affairs Palo Alto Health Care System. Both subject-matter literature and expert-opinions suggest that the pre-operative HbA1c level can be highly associated with post-operative complications. Hence, it is important to have a good understanding on surgical patient's pre-operative HbA1c level (e.g., their mean estimate) for conducting necessary post-operative interventional care. Additional information about the patient selection and clinical background can be found in¹⁸.

The dataset consists of 247 patients who underwent carotid revascularization procedures received pre- and post-operative MRI scans at VA Palo Alto Health Care System from 2004 to 2011. Of the 247 patients, we only use 180 patients with completed neuropsychological testing for the analysis. 50 of the 180 patients had a missing preoperative HbA1c value. We conduct some exploratory analyses to evaluate the presence of MNAR. Specifically, we fit a selection model via a two-stage approach to the data. Table 5 shows the results. After controlling for diabetes status, older patients have a significantly lower HbA1c level than younger patients with a p -value of 0.03. After controlling for age, diabetic patients have significantly higher HbA1c values than non-diabetic patients with a p -value of < 0.0001 . After controlling for the diabetes status, male patients are more likely to have their HbA1c observed than female patients with a p -value of < 0.0001 . After controlling for sex, diabetic patients are more likely to have their HbA1c observed than non-diabetic patients with a p -value of < 0.0001 . More importantly, the fitted selection model also produces a highly significant ρ estimate of 0.99 with a 95% CI of (0.97, 1.00), showing some evidence that the missingness is MNAR after controlling for these covariates. That is, patients with lower HbA1c levels are more likely to be missing after controlling for the covariates.

Finally, we apply the sensitivity analysis approach to estimate the mean of HbA1c, with ρ ranging from -1 to +1 and 5 imputations for each ρ . For both PMM-MI and NNMI, age and diabetic status are used to select imputing sets for each missing HbA1c observation. For NNMI, age is categorized into four groups based on its quartiles. Figure 1 and Table 6 provide the results for estimation of mean HbA1c level. The CC analysis, assuming MCAR, produces a mean estimate of 6.497, which is likely to over-estimate the mean HbA1c level based on the suspicion of MNAR. Note that the over-estimation HbA1c level in CC could lead to unnecessary and costly post-operative intervention for complications due to surgery. The PMM-MI method produces a mean estimate of 4.692, which is likely to over-correct the over-estimation of CC since it is mainly driven by the diabetic status. On the other hand, NNMI, which accounts for potential MNAR through incorporating the information from various ρ 's, produces a mean HbA1c level ranging from 2.45% ($\rho = -1$) higher than that of CC to 5.08% ($\rho = 1$) lower than that of CC. NNMI changes from producing a mean HbA1c level higher than CC to lower than CC as ρ becomes greater than -0.60 (Figure 1). Table 7 provides regression analysis for evaluating whether the diabetic status and age are predictive of the pre-operative HbA1c level for carotid patients. CC analysis indicates both diabetic status and age are predictive of the pre-operative HbA1c level. NNMI reaches the same conclusion as CC with regardless of the ρ value. In contrast, PMM-MI indicates only the diabetic status is predictive of the pre-operative HbA1c level, which is likely due to the over-correction of over-estimation of CC analysis as mentioned earlier.

5 | DISCUSSION

In this paper we propose a multiple imputation-based sensitivity analysis approach to handle missing data subject to MNAR under a selection model framework. The proposed approach does not directly use a model to perform imputation. Instead, the model is only utilized to first conceptualize the potential relationship between the variable subject to missingness and the missingness probability (i.e. the potential missingness mechanism) and then determine the sensitivity parameter. Specifically, we use Heckman's selection model to describe the missingness mechanism, in which the missingness mechanism is determined through the correlation coefficient ρ between the variable subject to missingness and a latent variable associated with the selection probability. The correlation coefficient ρ controls the missingness mechanism and, therefore, can be used as a sensitivity parameter. In our approach, the sensitivity parameter ρ is only employed to define imputing sets from the observed data, and produce multiple imputations for the missing cases. In this regard, the proposed approach can be considered as a nonparametric multiple imputation approach and should be more robust to misspecification of the selection model than the sensitivity analysis approaches directly used the selection model to perform imputation. In addition, the sensitivity parameter pre-specified in the analysis is standardized and indicates the magnitude and direction of MNAR.

To demonstrate that the proposed approach can be used as a sensitivity analysis approach for data subject to MNAR, we first consider a simple scenario in which there is no fully observed covariate. This scenario does not allow one to use Heckman's selection model to perform imputation since the estimation of Heckman's selection model requires additional fully observed covariates. Therefore, we do not compare our approach with any sensitivity analysis approaches directly used Heckman's selection model to perform imputation in this paper. Based on the simulation study results, the proposed approach can correct some bias of the CC analysis even if the pre-specified sensitivity parameter deviates from the true value. When the data are not from the selection model, our limited exploratory simulation study suggests that the Spearman correlation coefficient might be a better sensitivity analysis parameter than Pearson correlation coefficient and Kendall's τ . We then generalize the sensitivity analysis approach for a situation with additional fully observed covariates, which are used to select the subjects with their outcome values observed similar to a subject with a missing outcome to define an imputing set for the subject with a missing outcome, while estimating the regression coefficients for the relationship between the missing outcome and the fully observed covariates. We compare the proposed sensitivity analysis approach with the CC, which is under the MCAR assumption, and PMM-MI, which is under the MAR assumption, methods. Based on the simulation study results, the proposed approach can correct some bias of the CC and PMM-MI analyses when the sensitivity parameter is correctly specified.

Even if this paper only focuses on developing a sensitivity analysis approach for data subject to MNAR, in which the sensitivity analysis parameter, i.e. the correlation coefficient between the missing variable and the selection probability, is specified, the proposed approach can be easily modified to incorporate the existing fully observed covariates into Heckman's selection model to estimate the correlation coefficient, similar to the extensions of delta-adjustment procedures that have been proposed^{19,20}. Specifically, with covariates the correlation coefficient ρ in the selection models can be estimated, although the estimation process can be complicated and unstable²¹. The information from covariates can be used to derive two predictive scores to identify imputing sets for missing observations, following the nearest neighbor-based multiple imputation idea in the case of MAR^{7,8}. This will allow us to compare the proposed approach with the approaches that directly uses Heckman's selection model to perform imputation^{14,15} and will be explored in the future study.

The proposed sensitivity analysis approach is specifically developed for MNAR induced through a selection model framework and is not expected to perform well if MNAR is not induced through a selection model framework. For example, if MNAR is induced by a pattern mixture model, the correlation coefficient between the missing variable and the selection probability can be low but the magnitude of MNAR is high. In this scenario, the proposed non-parametric multiple imputation procedures will not be able to correct the bias of the CC analysis. As a result, this non-parametric multiple imputation-based sensitivity analysis approach is not applicable to be used to perform sensitivity analysis for data subject to MNAR induced by a pattern-mixture model framework. In addition, the proposed approach cannot be generalized to perform sensitivity analysis for a situation with

a missing binary outcome subject to MNAR since a binary outcome only has a value of 0 or 1 and, therefore, cannot be directly used to define a nearest neighborhood for a missing binary outcome observation.

6 | ACKNOWLEDGEMENTS

Dr. Hsu's work was partially supported by National Institutes of Health grant P30 CA023074. Dr. Zhou's work was partially supported by National Institutes of Health grant R01NS070308.

References

1. Rubin DB. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J Am Stat Assoc.* 1977; **72(359)**:538–543.
2. Little RJA. Pattern-Mixture Models for Multivariate Incomplete Data. *Journal of the American Statistical Association* 1993; **88**: 125–134.
3. Yan X, Lee S, Li N. Missing Data Handling Methods in Medical Device Clinical Trials. *Journal of Biopharmaceutical Statistics* 2009; **19(6)**: 1085–1098.
4. Heckman JJ. Shadow Prices, Market Wages, and Labor Supply. *Econometrica* 1974; **42**: 679–694.
5. Heckman JJ. Sample Selection Bias as a Specification Error. *Econometrica* 1979; **47**: 153–161.
6. Copas, JB, Li HG. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1997; **59**: 55–95.
7. Long Q, Hsu CH, Li Y. Doubly Robust Nonparametric Multiple Imputation for Ignorable Missing Data. *Statistica Sinica* 2012; **22**: 149–172.
8. Hsu CH, He Y, Li Y, Long Q, Friese R. Doubly robust multiple imputation using kernel-based techniques. *Biometrical Journal* 2016; **58**: 588–606.
9. Schenker N, Taylor JMG. Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis* 1996; **22**: 425–446.
10. Siddique J, Belin TR. Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine* 2008; **27**: 83–102.

11. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987. ISBN 9780471087052.
12. Van Buren S. *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall; 2012. ISBN 1439868247.
13. Yuan Y. *Sensitivity analysis in multiple imputation for missing data*. March 23-26, 2014. Washington, DC. SAS Global Forum SAS Institute Inc.
14. Galimard JE, Chevret S, Protopopescu C, Resche-Rigon M. A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine* 2016; **35**: 2907–2920.
15. Ogundimu EO, Collins GS. A robust imputation method for missing responses and covariates in sample selection models. *Statistical Methods in Medical Research* 2019; **28**: 102–116.
16. Jolani S. Dual Imputation Strategies for Analyzing Incomplete Data. *Dissertation. University of Utrecht*, Dec 7, 2012.
17. Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: John Wiley & Sons; 2002. ISBN 9780471183860.
18. Zhou W, Hitchner E, Gillis K, Sun L, Floyd R, Lane B, Rosen A. Prospective neurocognitive evaluation of patients undergoing carotid interventions. *Journal of Vascular Surgery* 2012; **56**: 1571–1578.
19. Moreno-Betancur M, Chavance M. Sensitivity analysis of incomplete longitudinal data departing from the missing at random assumption: methodology and application in a clinical trial with drop-outs. *Stat Methods Med Res.* 2013; **25**: 1471–1489.
20. Liublinska V, Rubin DB. Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat Med.* 2014; **33**: 4170–4185.
21. Leung SF, Yu S. Collinearity and two-step estimation of sample selection models: problems, origins and remedies. *Comput Econ.* 2000; **15**: 173–199.

TABLE 1 Monte Carlo results for MNAR from a selection model framework, where replications=500, $N = 200$, $MI=5$, $\mu_y = 1.5$, $\mu_x = 0.5$ and missingness rate=0.31.

Method	Bias ^a	SD ^b	SE ^c	CR ^d	Bias	SD	SE	CR	Bias	SD	SE	CR
$\sigma_y = 0.8$				$\sigma_y = 1.0$				$\sigma_y = 1.2$				
$\rho(x, y) = -0.50$												
FO	0.002	0.060	0.056	0.938	0.002	0.075	0.071	0.942	0.003	0.090	0.085	0.938
CC	-0.200	0.064	0.063	0.128	-0.248	0.081	0.079	0.124	-0.297	0.097	0.095	0.134
NNMI	-0.019	0.071	0.062	0.908	-0.023	0.089	0.078	0.902	-0.027	0.107	0.094	0.904
$\rho(x, y) = -0.30$												
FO	0.002	0.060	0.057	0.942	0.002	0.075	0.071	0.942	0.003	0.089	0.085	0.942
CC	-0.119	0.068	0.066	0.580	-0.148	0.085	0.083	0.594	-0.177	0.101	0.099	0.600
NNMI	-0.002	0.072	0.066	0.926	-0.002	0.090	0.083	0.936	-0.001	0.107	0.100	0.930
$\rho(x, y) = -0.10$												
FO	0.002	0.059	0.057	0.936	0.003	0.074	0.071	0.936	0.003	0.089	0.085	0.936
CC	-0.040	0.069	0.068	0.882	-0.049	0.087	0.085	0.888	-0.059	0.104	0.102	0.884
NNMI	0.004	0.072	0.073	0.954	0.005	0.090	0.091	0.954	0.006	0.108	0.109	0.954
$\rho(x, y) = 0.00$												
FO	0.002	0.059	0.057	0.934	0.003	0.074	0.071	0.934	0.003	0.089	0.085	0.934
CC	0.002	0.068	0.068	0.952	0.002	0.086	0.085	0.952	0.003	0.103	0.102	0.952
NNMI	0.002	0.073	0.075	0.962	0.003	0.091	0.094	0.962	0.004	0.109	0.113	0.962
$\rho(x, y) = 0.10$												
FO	0.002	0.059	0.057	0.932	0.003	0.074	0.071	0.934	0.003	0.088	0.085	0.934
CC	0.041	0.068	0.068	0.916	0.052	0.085	0.085	0.912	0.062	0.103	0.102	0.908
NNMI	-0.002	0.071	0.072	0.958	-0.002	0.088	0.090	0.956	-0.003	0.106	0.108	0.960
$\rho(x, y) = 0.30$												
FO	0.002	0.058	0.057	0.940	0.003	0.073	0.071	0.940	0.003	0.088	0.085	0.940
CC	0.123	0.069	0.066	0.528	0.154	0.086	0.083	0.528	0.185	0.104	0.099	0.538
NNMI	0.008	0.071	0.066	0.922	0.010	0.089	0.083	0.914	0.012	0.108	0.099	0.926
$\rho(x, y) = 0.50$												
FO	0.002	0.057	0.056	0.948	0.003	0.072	0.071	0.948	0.004	0.087	0.085	0.948
CC	0.204	0.066	0.063	0.122	0.256	0.082	0.079	0.116	0.307	0.099	0.095	0.112
NNMI	0.024	0.070	0.063	0.902	0.031	0.088	0.079	0.896	0.037	0.106	0.094	0.900

^aAverage of 500 point estimates.

^bEmpirical standard deviation.

^cAverage estimated standard error.

^dCoverage rate of 500 95% confidence intervals.



TABLE 2 Monte Carlo results with misspecified ρ for selecting imputing sets, where replications=500, $N = 200$, $MI=5$, $\mu_y = 1.5$, $\mu_x = 0.5$ and missingness rate=0.31.

Method	Bias ^a	SD ^b	SE ^c	CR ^d	Bias	SD	SE	CR	Bias	SD	SE	CR
	$\sigma_y = 0.8$				$\sigma_y = 1.0$				$\sigma_y = 1.2$			
true $\rho(x, y) = -0.30$												
FO	0.002	0.060	0.057	0.942	0.002	0.075	0.071	0.942	0.003	0.089	0.085	0.942
CC	-0.119	0.068	0.066	0.580	-0.148	0.085	0.083	0.594	-0.177	0.101	0.099	0.600
NNMI($\rho^e=-0.55$)	0.089	0.076	0.067	0.722	0.110	0.094	0.084	0.740	0.133	0.112	0.100	0.728
NNMI($\rho=-0.50$)	0.069	0.075	0.067	0.814	0.087	0.093	0.083	0.832	0.104	0.110	0.100	0.828
NNMI($\rho=-0.45$)	0.050	0.075	0.066	0.872	0.064	0.094	0.083	0.866	0.077	0.110	0.099	0.872
NNMI($\rho=-0.40$)	0.032	0.074	0.065	0.904	0.040	0.092	0.082	0.906	0.048	0.108	0.099	0.914
NNMI($\rho=-0.35$)	0.014	0.073	0.066	0.932	0.017	0.090	0.083	0.928	0.021	0.108	0.100	0.928
NNMI($\rho=-0.30$)	-0.003	0.073	0.067	0.918	-0.004	0.089	0.084	0.920	-0.005	0.106	0.100	0.920
NNMI($\rho=-0.25$)	-0.021	0.072	0.068	0.912	-0.027	0.088	0.084	0.912	-0.032	0.106	0.101	0.916
NNMI($\rho=-0.20$)	-0.039	0.071	0.069	0.888	-0.049	0.087	0.086	0.888	-0.057	0.106	0.103	0.898
NNMI($\rho=-0.15$)	-0.057	0.072	0.070	0.854	-0.072	0.089	0.087	0.854	-0.086	0.105	0.104	0.870
NNMI($\rho=-0.10$)	-0.076	0.072	0.071	0.806	-0.095	0.088	0.090	0.830	-0.114	0.104	0.107	0.830
NNMI($\rho=-0.05$)	-0.097	0.070	0.073	0.760	-0.120	0.088	0.091	0.774	-0.145	0.104	0.110	0.778
true $\rho(x, y) = 0.30$												
FO	0.002	0.058	0.057	0.940	0.003	0.073	0.071	0.940	0.003	0.088	0.085	0.940
CC	0.123	0.069	0.066	0.528	0.154	0.086	0.083	0.528	0.185	0.104	0.099	0.538
NNMI($\rho=0.05$)	0.101	0.071	0.073	0.728	0.127	0.089	0.092	0.742	0.153	0.108	0.110	0.748
NNMI($\rho=0.10$)	0.079	0.070	0.071	0.814	0.101	0.089	0.090	0.830	0.120	0.106	0.107	0.816
NNMI($\rho=0.15$)	0.061	0.071	0.071	0.868	0.077	0.087	0.088	0.880	0.092	0.106	0.105	0.876
NNMI($\rho=0.20$)	0.043	0.070	0.068	0.906	0.053	0.088	0.086	0.918	0.063	0.106	0.103	0.906
NNMI($\rho=0.25$)	0.023	0.070	0.067	0.930	0.029	0.087	0.084	0.928	0.035	0.106	0.100	0.928
NNMI($\rho=0.30$)	0.006	0.070	0.067	0.940	0.009	0.088	0.083	0.932	0.010	0.106	0.101	0.924
NNMI($\rho=0.35$)	-0.011	0.070	0.066	0.934	-0.012	0.087	0.083	0.934	-0.015	0.106	0.100	0.924
NNMI($\rho=0.40$)	-0.029	0.071	0.066	0.914	-0.036	0.088	0.084	0.922	-0.043	0.107	0.100	0.914
NNMI($\rho=0.45$)	-0.047	0.071	0.067	0.890	-0.058	0.089	0.083	0.884	-0.069	0.108	0.100	0.878
NNMI($\rho=0.50$)	-0.065	0.072	0.066	0.828	-0.082	0.090	0.083	0.822	-0.098	0.108	0.099	0.830
NNMI($\rho=0.55$)	-0.084	0.072	0.066	0.774	-0.106	0.090	0.083	0.750	-0.127	0.110	0.100	0.768

^aAverage of 500 point estimates.

^bEmpirical standard deviation.

^cAverage estimated standard error.

^dCoverage rate of 500 95% confidence intervals.

^e ρ used to generate random draws of ρ for selecting imputing sets.

TABLE 3 Monte Carlo results with data generated from exponential distribution, where replications=500, $N = 200$, $MI=5$, $\lambda_y = 1.0$, $\lambda_x = 1.0$ and missingness rate=0.30, i.e. $Pr(X > 0.357)$.

Method	Bias ^a	SD ^b	SE ^c	CR ^d	Bias	SD	SE	CR
$r^e = -0.37; s^f = -0.48; \tau^g = -0.33$					$r = 0.45; s = 0.48; \tau = 0.33$			
FO	0.002	0.075	0.070	0.926	0.003	0.073	0.071	0.940
CC	-0.237	0.066	0.065	0.078	0.205	0.095	0.090	0.352
NNMI _{$\rho=r$}	-0.132	0.075	0.070	0.560	-0.007	0.078	0.074	0.920
NNMI _{$\rho=s$}	-0.089	0.080	0.073	0.728	-0.014	0.075	0.072	0.918
NNMI _{$\rho=\tau$}	-0.146	0.069	0.069	0.492	0.031	0.078	0.072	0.916
$r = -0.23; s = -0.29; \tau = -0.19$					$r = 0.26; s = 0.29; \tau = 0.19;$			
FO	0.003	0.074	0.070	0.926	0.003	0.073	0.070	0.936
CC	-0.137	0.077	0.074	0.538	0.129	0.094	0.089	0.716
NNMI _{$\rho=r$}	-0.070	0.082	0.078	0.802	-0.017	0.081	0.077	0.916
NNMI _{$\rho=s$}	-0.051	0.083	0.079	0.828	-0.025	0.076	0.073	0.908
NNMI _{$\rho=\tau$}	-0.083	0.081	0.077	0.748	0.009	0.080	0.074	0.930
$r = -0.08; s = -0.10; \tau = -0.06$					$r = 0.08; s = 0.10; \tau = 0.06$			
FO	0.003	0.074	0.070	0.930	0.003	0.074	0.070	0.930
CC	-0.044	0.085	0.081	0.858	0.046	0.090	0.086	0.930
NNMI _{$\rho=r$}	-0.027	0.084	0.083	0.898	-0.015	0.084	0.083	0.922
NNMI _{$\rho=s$}	-0.021	0.083	0.083	0.904	-0.020	0.080	0.081	0.922
NNMI _{$\rho=\tau$}	-0.030	0.084	0.082	0.882	-0.004	0.081	0.080	0.928
$r = 0.00; s = 0.00; \tau = 0.00$								
FO	0.003	0.074	0.070	0.932				
CC	0.003	0.088	0.084	0.926				
NNMI _{$\rho=r$}	-0.015	0.085	0.086	0.934				
NNMI _{$\rho=s$}	-0.014	0.082	0.084	0.936				
NNMI _{$\rho=\tau$}	-0.011	0.083	0.083	0.934				

^aAverage of 500 point estimates.

^bEmpirical standard deviation.

^cAverage estimated standard error.

^dCoverage rate of 500 95% confidence intervals.

^ePearson correlation coefficient.

^fSpearman correlation coefficient.

^gKendall's τ .

TABLE 4 Monte Carlo results: Regression analysis, where replications=500, $N = 300$, $MI=5$, and missingness rate=0.39.

Method	Est ^a	SD ^b	SE ^c	CR ^d	Est	SD	SE	CR	Est	SD	SE	CR
$\beta_0 = 0.50$				$\beta_1 = 2.0$				$\beta_2 = 2.0$				
true $\rho(x, y) = -0.50$												
FO	0.510	0.127	0.130	95.8	1.992	0.113	0.116	94.8	1.992	0.193	0.201	94.8
CC	0.191	0.157	0.155	49.0	2.136	0.146	0.140	81.0	1.859	0.229	0.241	93.2
PMM-MI	0.202	0.167	0.156	55.2	2.133	0.155	0.142	83.0	1.841	0.253	0.244	89.8
NNMI ₃	0.493	0.168	0.149	90.6	2.023	0.159	0.134	90.2	1.885	0.252	0.231	90.4
NNMI ₄	0.462	0.170	0.149	89.8	2.033	0.160	0.134	89.0	1.916	0.254	0.233	91.8
NNMI ₅	0.445	0.169	0.149	90.2	2.039	0.158	0.134	88.8	1.927	0.255	0.235	91.2
NNMI ₆	0.433	0.167	0.151	89.4	2.048	0.158	0.137	88.4	1.923	0.255	0.237	91.8
NNMI ₈	0.411	0.170	0.151	87.6	2.065	0.160	0.138	88.8	1.907	0.260	0.242	91.0
NNMI ₁₀	0.387	0.170	0.154	86.6	2.083	0.159	0.140	87.2	1.882	0.255	0.247	90.4
true $\rho(x, y) = -0.50$												
FO	0.510	0.127	0.130	95.6	1.993	0.113	0.116	94.8	1.991	0.192	0.201	94.6
CC	0.319	0.160	0.162	81.4	2.076	0.151	0.146	89.6	1.916	0.235	0.252	94.8
PMM-MI	0.325	0.169	0.163	81.6	2.075	0.162	0.149	90.8	1.902	0.256	0.259	92.2
NNMI ₃	0.526	0.170	0.157	92.8	2.003	0.162	0.142	91.2	1.901	0.246	0.241	92.4
NNMI ₄	0.500	0.171	0.158	92.4	2.012	0.162	0.143	92.0	1.931	0.253	0.245	93.0
NNMI ₅	0.485	0.172	0.159	92.4	2.016	0.160	0.144	92.0	1.940	0.254	0.245	93.8
NNMI ₆	0.474	0.173	0.159	91.8	2.023	0.160	0.143	91.8	1.940	0.257	0.246	94.0
NNMI ₈	0.453	0.171	0.161	92.0	2.036	0.162	0.147	92.2	1.937	0.258	0.253	94.6
NNMI ₁₀	0.434	0.170	0.162	92.0	2.050	0.163	0.148	90.0	1.924	0.259	0.256	93.8
true $\rho(x, y) = -0.10$												
FO	0.510	0.128	0.130	95.4	1.994	0.114	0.116	95.2	1.991	0.192	0.201	94.6
CC	0.449	0.163	0.166	95.0	2.017	0.152	0.150	94.6	1.965	0.242	0.258	96.4
PMM-MI	0.454	0.170	0.169	92.6	2.014	0.159	0.154	94.6	1.956	0.262	0.266	94.2
NNMI ₃	0.542	0.168	0.168	93.6	2.00	0.158	0.150	94.2	1.895	0.245	0.258	92.4
NNMI ₄	0.519	0.169	0.170	93.8	2.005	0.159	0.152	94.2	1.930	0.248	0.257	95.0
NNMI ₅	0.503	0.170	0.172	94.2	2.010	0.158	0.152	95.0	1.945	0.248	0.259	94.8
NNMI ₆	0.497	0.170	0.170	94.0	2.011	0.159	0.152	94.8	1.947	0.254	0.265	94.6
NNMI ₈	0.479	0.172	0.169	94.6	2.019	0.159	0.153	94.0	1.955	0.253	0.264	95.6
NNMI ₁₀	0.469	0.172	0.169	94.0	2.025	0.162	0.152	93.2	1.951	0.261	0.263	93.0
true $\rho(x, y) = 0.0$												
FO	0.510	0.128	0.130	95.8	1.994	0.114	0.116	95.4	1.991	0.191	0.201	94.8
CC	0.509	0.164	0.167	95.4	1.991	0.153	0.150	95.4	1.993	0.246	0.259	96.2
PMM-MI	0.517	0.176	0.167	94.4	1.986	0.165	0.153	93.6	1.985	0.264	0.262	93.4
NNMI ₃	0.546	0.173	0.171	94.4	1.996	0.165	0.154	92.8	1.905	0.248	0.264	94.4
NNMI ₄	0.528	0.172	0.171	94.2	1.995	0.163	0.153	93.8	1.946	0.252	0.266	96.4
NNMI ₅	0.519	0.171	0.170	94.6	1.994	0.161	0.154	95.2	1.965	0.256	0.265	95.0
NNMI ₆	0.517	0.174	0.171	95.0	1.991	0.162	0.154	94.8	1.975	0.259	0.269	95.4
NNMI ₈	0.513	0.174	0.168	93.4	1.995	0.164	0.153	93.6	1.981	0.261	0.268	95.0
NNMI ₁₀	0.510	0.175	0.169	94.2	1.997	0.165	0.153	93.4	1.977	0.263	0.269	95.8
true $\rho(x, y) = 0.10$												
FO	0.510	0.128	0.130	95.4	1.995	0.114	0.116	95.2	1.990	0.192	0.201	95.0
CC	0.571	0.165	0.166	93.0	1.964	0.154	0.150	94.2	2.024	0.250	0.258	96.0
PMM-MI	0.579	0.173	0.166	91.8	1.964	0.162	0.152	93.2	2.007	0.269	0.261	94.0
NNMI ₃	0.554	0.175	0.169	92.6	1.988	0.165	0.151	92.8	1.910	0.258	0.255	93.6
NNMI ₄	0.542	0.174	0.169	93.8	1.985	0.162	0.152	93.6	1.955	0.264	0.258	92.4
NNMI ₅	0.537	0.174	0.172	94.0	1.980	0.163	0.150	92.0	1.983	0.264	0.261	94.6
NNMI ₆	0.541	0.172	0.171	95.0	1.975	0.163	0.153	93.0	1.996	0.265	0.262	94.6
NNMI ₈	0.543	0.178	0.170	93.0	1.972	0.164	0.152	92.2	2.011	0.276	0.264	93.6
NNMI ₁₀	0.546	0.178	0.169	91.8	1.972	0.164	0.153	93.6	2.013	0.274	0.264	94.0
true $\rho(x, y) = 0.30$												
FO	0.509	0.129	0.130	94.8	1.996	0.114	0.116	95.0	1.990	0.192	0.201	95.6
CC	0.700	0.160	0.162	76.6	1.910	0.150	0.146	89.8	2.073	0.244	0.252	94.2
PMM-MI	0.707	0.166	0.162	73.6	1.907	0.160	0.148	89.0	2.062	0.257	0.257	93.4
NNMI ₃	0.574	0.168	0.158	90.4	1.986	0.164	0.141	91.8	1.905	0.248	0.241	92.4
NNMI ₄	0.558	0.169	0.157	92.4	1.979	0.164	0.141	91.4	1.963	0.253	0.241	91.8
NNMI ₅	0.560	0.170	0.158	91.8	1.971	0.161	0.141	91.0	1.987	0.254	0.246	94.0
NNMI ₆	0.568	0.171	0.159	90.0	1.966	0.163	0.142	90.4	1.998	0.259	0.247	93.2
NNMI ₈	0.575	0.174	0.159	88.2	1.957	0.164	0.144	89.4	2.025	0.263	0.251	92.8
NNMI ₁₀	0.590	0.171	0.161	89.4	1.945	0.164	0.145	90.4	2.032	0.263	0.255	92.8
true $\rho(x, y) = 0.50$												
FO	0.509	0.130	0.130	94.6	1.997	0.114	0.116	95.2	1.990	0.193	0.201	95.6
CC	0.827	0.153	0.155	42.4	1.854	0.140	0.139	81.4	2.124	0.236	0.240	92.6
PMM-MI	0.835	0.161	0.157	46.4	1.847	0.148	0.142	82.8	2.116	0.253	0.247	91.4
NNMI ₃	0.603	0.164	0.150	86.2	1.973	0.159	0.136	90.8	1.914	0.242	0.232	91.4
NNMI ₄	0.594	0.169	0.150	88.2	1.965	0.158	0.135	90.6	1.968	0.253	0.230	91.8
NNMI ₅	0.599	0.171	0.150	86.4	1.957	0.155	0.136	90.0	1.991	0.258	0.232	91.8
NNMI ₆	0.601	0.170	0.151	85.4	1.950	0.156	0.136	90.6	2.012	0.264	0.233	92.2
NNMI ₈	0.616	0.168	0.151	85.8	1.938	0.156	0.136	89.6	2.034	0.261	0.238	91.8
NNMI ₁₀	0.634	0.169	0.153	81.8	1.926	0.156	0.140	88.6	2.033	0.262	0.241	93.4

^a Average of 500 point estimates.
^b Empirical standard deviation.
^c Average estimated standard error.
^d Coverage rate of 500 95% confidence intervals.

TABLE 5 Data Analysis: results of the fitted sample selection model

HbA1c Value Model		
Variable	coeff \pm SE	p-value
Intercept	6.541 \pm 0.638	<0.0001
Age	-0.019 \pm 0.009	0.03
Diabetes	1.773 \pm 0.196	<0.0001
σ_y	1.218 \pm 0.085	<0.0001
Selection Probability Model		
Intercept	-7.620 \pm 0.059	< 0.0001
Male	7.827 \pm 0.059	<0.0001
Diabetes	0.897 \pm 0.189	<0.0001
ρ	0.987 \pm 0.010	<0.0001

TABLE 6 Sensitivity analysis for estimation of mean HbA1c level

Method	Est ^a	SE ^b	95% CI ^c
CC	6.497	0.113	(6.276, 6.718)
PMM-MI	4.692	0.232	(4.237, 5.147)
NNMI($\rho = -1.00$)	6.656	0.126	(6.400, 6.913)
NNMI($\rho = -0.90$)	6.590	0.136	(6.306, 6.875)
NNMI($\rho = -0.80$)	6.584	0.105	(6.377, 6.791)
NNMI($\rho = -0.70$)	6.542	0.101	(6.342, 6.741)
NNMI($\rho = -0.60$)	6.517	0.101	(6.317, 6.716)
NNMI($\rho = -0.50$)	6.494	0.107	(6.282, 6.707)
NNMI($\rho = -0.40$)	6.458	0.097	(6.267, 6.648)
NNMI($\rho = -0.30$)	6.437	0.091	(6.258, 6.616)
NNMI($\rho = -0.20$)	6.413	0.096	(6.225, 6.602)
NNMI($\rho = -0.10$)	6.386	0.091	(6.208, 6.564)
NNMI($\rho = 0.00$)	6.381	0.093	(6.199, 6.564)
NNMI($\rho = 0.10$)	6.354	0.090	(6.179, 6.530)
NNMI($\rho = 0.20$)	6.331	0.090	(6.154, 6.508)
NNMI($\rho = 0.30$)	6.318	0.088	(6.145, 6.491)
NNMI($\rho = 0.40$)	6.299	0.089	(6.125, 6.473)
NNMI($\rho = 0.50$)	6.282	0.089	(6.108, 6.456)
NNMI($\rho = 0.60$)	6.263	0.089	(6.088, 6.438)
NNMI($\rho = 0.70$)	6.247	0.093	(6.065, 6.430)
NNMI($\rho = 0.80$)	6.218	0.091	(6.040, 6.395)
NNMI($\rho = 0.90$)	6.174	0.105	(5.966, 6.382)
NNMI($\rho = 1.00$)	6.167	0.097	(5.976, 6.358)

^aPoint estimate of the average HbA1c level.

^bStandard error.

^c95% confidence interval.

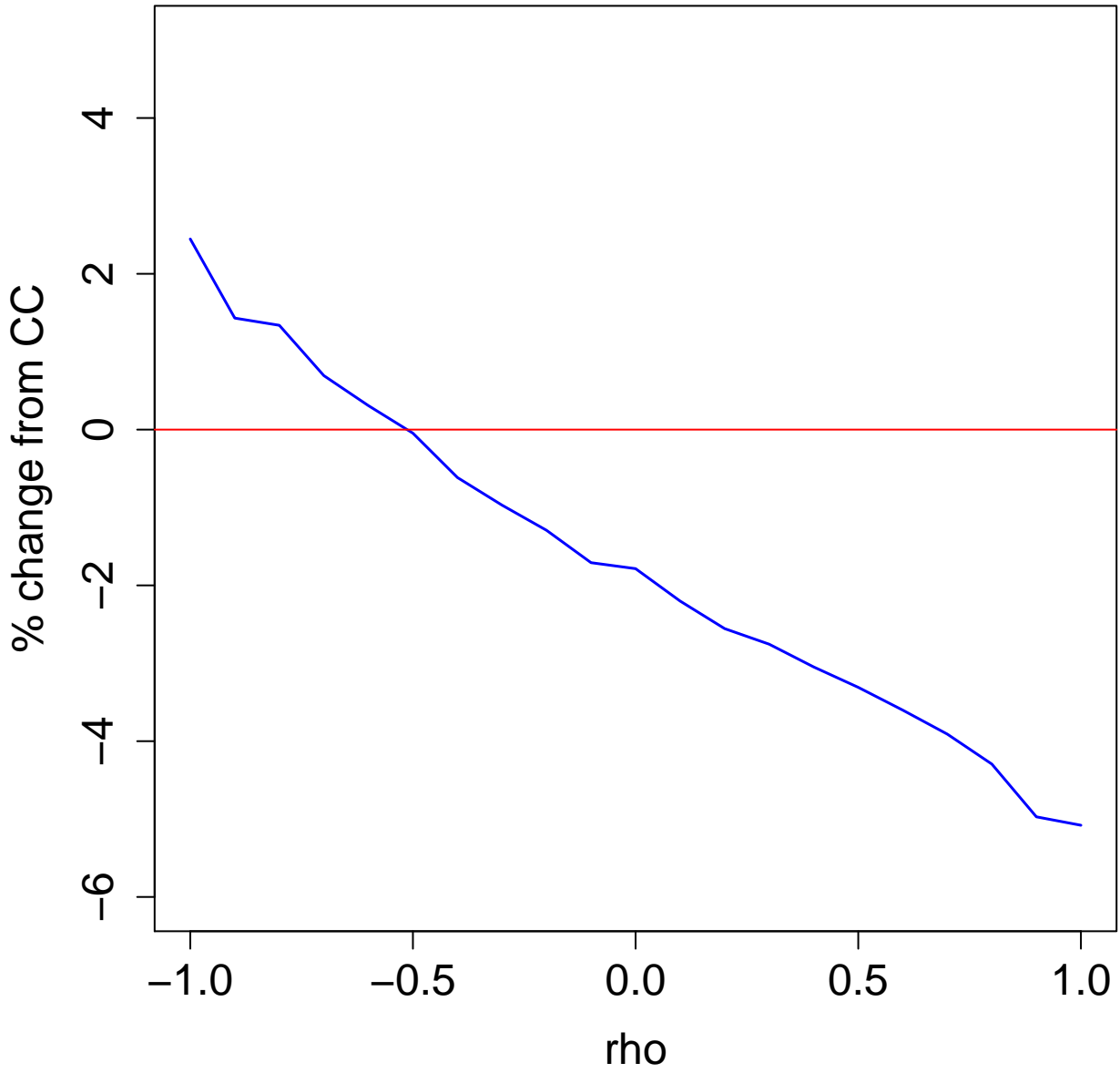


FIGURE 1 NNMI % changes from CC analysis

TABLE 7 Sensitivity analysis for regression analysis of HbA1c data

Method	Intercept			Diabetes			Age		
	Est ^a	SE ^b	p-value ^c	Est	SE	p-value	Est	SE	p-value
CC	7.326	0.779	< 0.0001	1.611	0.173	< 0.0001	-0.023	0.011	0.04
PMM-MI	5.816	2.030	< 0.01	2.258	0.444	< 0.0001	-0.029	0.029	0.31
NNMI($\rho = -1.00$)	8.343	0.944	< 0.0001	1.604	0.203	< 0.0001	-0.033	0.013	0.01
NNMI($\rho = -0.90$)	7.953	1.044	< 0.0001	1.588	0.331	< 0.0001	-0.029	0.014	0.04
NNMI($\rho = -0.80$)	7.720	0.801	< 0.0001	1.620	0.184	< 0.0001	-0.026	0.011	0.02
NNMI($\rho = -0.70$)	7.522	0.748	< 0.0001	1.616	0.171	< 0.0001	-0.023	0.010	0.02
NNMI($\rho = -0.60$)	7.454	0.715	< 0.0001	1.597	0.169	< 0.0001	-0.023	0.010	0.02
NNMI($\rho = -0.50$)	7.368	0.697	< 0.0001	1.575	0.156	< 0.0001	-0.022	0.010	0.03
NNMI($\rho = -0.40$)	7.291	0.687	< 0.0001	1.580	0.162	< 0.0001	-0.021	0.010	0.03
NNMI($\rho = -0.30$)	7.261	0.679	< 0.0001	1.576	0.154	< 0.0001	-0.021	0.010	0.03
NNMI($\rho = -0.20$)	7.271	0.682	< 0.0001	1.568	0.157	< 0.0001	-0.021	0.010	0.03
NNMI($\rho = -0.10$)	7.269	0.703	< 0.0001	1.569	0.153	< 0.0001	-0.021	0.010	0.03
NNMI($\rho = 0.00$)	7.225	0.667	< 0.0001	1.558	0.158	< 0.0001	-0.021	0.009	0.02
NNMI($\rho = 0.10$)	7.170	0.660	< 0.0001	1.560	0.147	< 0.0001	-0.021	0.009	0.03
NNMI($\rho = 0.20$)	7.108	0.641	< 0.0001	1.553	0.148	< 0.0001	-0.020	0.009	0.03
NNMI($\rho = 0.30$)	7.071	0.632	< 0.0001	1.537	0.143	< 0.0001	-0.020	0.009	0.03
NNMI($\rho = 0.40$)	7.100	0.642	< 0.0001	1.540	0.140	< 0.0001	-0.020	0.009	0.02
NNMI($\rho = 0.50$)	7.087	0.639	< 0.0001	1.529	0.139	< 0.0001	-0.021	0.009	0.02
NNMI($\rho = 0.60$)	7.082	0.643	< 0.0001	1.532	0.141	< 0.0001	-0.021	0.009	0.02
NNMI($\rho = 0.70$)	7.098	0.655	< 0.0001	1.530	0.143	< 0.0001	-0.021	0.009	0.02
NNMI($\rho = 0.80$)	7.061	0.686	< 0.0001	1.524	0.151	< 0.0001	-0.021	0.010	0.03
NNMI($\rho = 0.90$)	7.174	0.819	< 0.0001	1.440	0.227	< 0.0001	-0.023	0.012	0.04
NNMI($\rho = 1.00$)	7.193	0.708	< 0.0001	1.520	0.150	< 0.0001	-0.024	0.010	0.02

^aRegression coefficient estimate.^bStandard error.^cp-value.