HHS Public Access

Author manuscript

Stat Methods Med Res. Author manuscript; available in PMC 2019 June 12.

Published in final edited form as:

Stat Methods Med Res. 2019 June ; 28(6): 1676–1688. doi:10.1177/0962280218772592.

Cox regression analysis with missing covariates via nonparametric multiple imputation

Chiu-Hsieh Hsu[1,2] and **Mandi Yu**[3]

[1]Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson, AZ, USA

[2]University of Arizona Cancer Center, University of Arizona, Tucson, AZ, USA

[3]Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Rockville, MD, USA

Abstract

We consider the situation of estimating Cox regression in which some covariates are subject to missing, and there exists additional information (including observed event time, censoring indicator and fully observed covariates) which may be predictive of the missing covariates. We propose to use two working regression models: one for predicting the missing covariates and the other for predicting the missing probabilities. For each missing covariate observation, these two working models are used to define a nearest neighbor imputing set. This set is then used to non-parametrically impute covariate values for the missing observation. Upon the completion of imputation, Cox regression is performed on the multiply imputed datasets to estimate the regression coefficients. In a simulation study, we compare the nonparametric multiple imputation approach with the augmented inverse probability weighted (AIPW) method, which directly incorporates the two working models into estimation of Cox regression, and the predictive mean matching imputation (PMM) method. We show that all approaches can reduce bias due to non-ignorable missing mechanism. The proposed nonparametric imputation method is robust to mis-specification of either one of the two working models and robust to mis-specification of the link function of the two working models. In contrast, the PMM method is sensitive to misspecification of the covariates included in imputation. The AIPW method is sensitive to the selection probability. We apply the approaches to a breast cancer dataset from Surveillance, Epidemiology and End Results (SEER) Program.

Keywords

Augmented inverse probability weighted method; Cox regression; missing covariates; multiple imputation; predictive mean matching

Corresponding author: Chiu-Hsieh Hsu, Department of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health and Arizona Cancer Center, University of Arizona, 1295 N Martin, PO Box 245211, Tucson 85724-5211, AZ, USA., pchhsu@email.arizona.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1 Introduction

For survival time data with covariates, Cox regression is often used to specify the relationship between survival time and covariates.[1] For time-independent covariates, Cox regression has the proportional hazards property. It estimates the regression coefficients of the model using the partial likelihood function without specifying the baseline hazard function.[2] The estimators of regression coefficients have been shown to be consistent, normally distributed and semi-parametrically efficient.[3] However, in many situations, some of the covariates are not fully observed. Missing covariates could compromise the asymptotic properties of the estimators if missing data are not accounted for in estimation. Specifically, it has been shown that the estimators of the regression coefficients derived from the subjects with all of the covariates observed (i.e. complete-case analysis) not only lose efficiency, but may also generate biased regression coefficient estimates when missingness depends on the survival outcome (i.e. survival time and censoring indicator).[4] When missingness depends on the survival outcome (i.e. survival time and censoring indicator) and some fully observed covariates, missing mechanism is considered as missing at random (MAR).[5] For the survival outcome data, MAR can be even further classified into two scenarios: failure- ignorable MAR (i.e. missingness does not depend on failure time) and censoring-ignorable MAR (i.e. missingness does not depend on censoring time but may depend on failure time).[6] When missingness is failure-ignorable MAR, complete-case analysis can still produce valid regression coefficient estimates. However, when missingness is censoring-ignorable MAR, complete-case analysis may produce biased regression coefficient estimates.

Several approaches have been proposed to deal with missing covariates in Cox regression. Of the existing approaches, the augmented inverse probability weighted (AIPW) method,[7,8] where the weight is derived from a fully specified model for the missing status conditional on the observed data and an augmentation term derived from a fully specified model for the missing covariate conditional on the observed data is added to estimation to correct the potential bias, has been shown to have a double robustness property. Specifically, the AIPW method uses two fully specified parametric models (one for the missing covariate and the other for the missing probability) to account for missing covariates while estimating the regression coefficients of Cox regression model. This indicates that at least one of the two models has to be correctly specified, including the distribution and link function for the missing covariate and the missing status, respectively. Of the two models, the model for the missing covariate is more important since in a sense it is directly associated with estimation in Cox regression. However, it is more challenging to correctly specify the model for the missing covariate than the model for the missing probability based on the observed data. Because of the double robustness property, the AIPW method is a popular method for researchers who do not want to solely rely on the model for estimating the conditional distribution of the missing covariate on the observed data in Cox regression with missing covariates. To weaken the reliance on parametric assumptions behind the two models, non-parametric regression has been used to estimate the two models without fully specifying the relationship between the missing covariates and the observed data.[9] As the dimensionality of the observed data increases, it becomes extremely difficult to use non- parametric regression

to estimate the two models. In addition, the AIPW method is also sensitive to misspecification of the missing probability model, because even mild lack of fit in outlying regions of the covariate space where the missing probability is extreme (i.e. very close to 1) translates into large errors in the weights.[5,10,11]

We previously developed a nonparametric multiple imputation (MI) approach to deal with missing data in a situation without censored data.[11] The approach indirectly uses two working models to recover information for missing data observations. Specifically, we use two working regression models, one for predicting the missing covariate values and one for predicting the missing probabilities. The parameter estimates from these two working models are then used to give two predictive scores for each subject, defined as the linear combination of the covariates in the corresponding model. The method then selects an imputing set of observations for each missing data observation, which consists of subjects who have their data fully observed and have similar predictive scores as the subject with missing data. Then the missing data value is randomly drawn from this imputing set. The idea is similar to predictive mean matching[12] and propensity score matching[13] in the missing data literature. In a situation with missing outcome data, we have shown that this nonparametric multiple imputation approach can generate a consistent mean estimator. In this paper, we generalize the nonparametric multiple imputation approach,[11] in which no statistical model is directly used to perform multiple imputation, to handle estimation of Cox regression with missing covariates to weaken the reliance on the two models and produce stable regression coefficient estimates even if the missing probability is extreme. Specifically, we propose to use two working regression models, one for predicting the missing covariates and one for predicting the missing probabilities, to derive two predictive scores to select an imputing set for each missing covariate observation. It has been shown that the survival outcome data (specifically cumulative baseline hazard and censoring indicator) need to be included in predicting the missing covariates.[14] In addition, the survival outcome data can be also included in the regression model for missing probabilities as the covariates to account for potentially censoring-ignorable MAR. The two working regression models are only used to derive two predictive scores to select an imputing set. Hence, the approach can easily handle the multi-dimensional structure of the observed data and is expected to be less affected by the mis-specification of the two working models (especially the mis-specification of the missing probability model) than the AIPW method. Due to the simplicity in estimation and the availability in statistical software, the MI method simply based on the predictive model of the missing covariates is widely used. Qi[15] compared the AIPW method with the MI method using predictive mean matching (PMM) based on multiple imputation by chain equations (MICE) in the estimation of Cox regression with missing covariates and concluded the PMM method is sensitive to misspecification of the predictive model of the missing covariates. In this paper, not only will we study the performance of the proposed multiple imputation approach but will also compare its performance with the AIPW and PMM methods.

This paper is organized as follows. In Section 2, we review the complete-case analysis and the AIPW method. In Section 3, we describe the proposed multiple imputation method and the associated properties. In Section 4, we apply the techniques to data from a breast cancer

study. In Section 5, we give results from a simulation study. A discussion follows in Section 6.

2 Review of methods

In this section, we begin with describing the setting of the situation: estimation of Cox regression with time- independent covariates and one of the covariates subject to missing. Let T denote the failure time, C denote the censoring time, $Y = min(T,C)$ denote the observed time, $\delta_t = I[T \leq C]$ denote the censoring indicator and $N(t)=\delta_t I(T \leq t)$ denote the counting process. Assume T has a hazard function of $\lambda(t) = \lambda_0(t)e^{\beta_x X+\beta_z Z}$ where $\lambda_0(t)$ is an unspecified baseline hazard function, X is subject to missing and Z is fully observed. Let d_X denote the missing indicator for X (i.e. δ_X =1 if X is observed; otherwise, 0) and π $Pr(\delta_X = 1)$ denote the selection probability. We assume that T and C are independent conditional on X and Z and X is missing at random (i.e. $E[\delta_X | Y, \delta_t, Z, X] = E[\delta_X | Y, \delta_t, Z]$) and there is a random sample of n subjects.

2.1 Complete-case analysis

The complete-case (CC) analysis of $\boldsymbol{\beta} = (\beta_X,\beta_Z)$ is based on the partial likelihood estimator using observations that have X observed. Let $r_i(\boldsymbol{\beta},t) = e^{\beta_x X+\beta_z Z_i} \equiv r_i^{(0)}(\boldsymbol{\beta},t)$ and $r_i^{(1)} = (X_i Z_i)^t r_i(\boldsymbol{\beta},t)$ The CC analysis involves solving the following estimating equations

$$U_{cc} = \sum_{i=1}^{n}\left[\delta_{t_i}\delta_{x_i}\left\{\begin{pmatrix}X_i\\Z_i\end{pmatrix} - \frac{S_{cc}^{(1)}(\boldsymbol{\beta},T_i)}{S_{cc}^{(0)}(\boldsymbol{\beta},T_i)}\right\}\right] = 0$$

where $S_{cc}^{(m)}(\boldsymbol{\beta},T_i) = n^{-1}\sum_{j=1}^{n}\delta_{x_i} I(T_j \geq T_i) r_j^{(m)}(\boldsymbol{\beta},T_i)$ for m = 0, 1. It is easy to implement the CC analysis and it is consistent when the missingness depends only on Z. However, it loses efficiency due to discarding data from incomplete observations, especially when the missing rate is greater than 25%,[16] and is inconsistent when missingness depends on T or δ_t.

2.2 AIPW method

The AIPW method was first proposed by Robins et al.[7] to modify the CC analysis to produce consistent estimators of $\boldsymbol{\beta}$ and furthermore improve efficiency of the CC analysis. The AIPW method has been studied and further developed by a few groups for various scenarios. For Cox regression with a missing covariate, it involves solving the estimating equations[8,9]

$$U_{AIPW} = \sum_{i=1}^{n}\left[\frac{\delta_{t_i}\delta_{x_i}}{\pi_i}\left\{\begin{pmatrix}X_i\\Z_i\end{pmatrix} - \frac{S_{AIPW}^{(1)}(\boldsymbol{\beta},T_i)}{S_{AIPW}^{(0)}(\boldsymbol{\beta},T_i)}\right\} + A_i(\boldsymbol{\beta},\pi_i)\right] = 0,$$

Where

$$S^{(m)}_{AIPW}(\boldsymbol{\beta}, T_i) = n^{-1}\sum_{j=1}^{n}\left\{\frac{\delta_{x_j}}{\pi_j} I(T_j \geq T_i) r_j^{(m)}(\boldsymbol{\beta}, T_i) + \left(1 - \frac{\delta_{x_j}}{\pi_j}\right) I(T_j \geq T_i) E\left[r_j^{(m)}(\boldsymbol{\beta}, T_i) | T_i, \delta_{t_i}, Z_i\right]\right\}$$

for m = 0,1 and

$$A_i(\boldsymbol{\beta}, \pi_i) = \left(1 - \frac{\delta_{x_j}}{\pi_j}\right) \int_0^{\tau} E\left[\begin{Bmatrix} X_i \\ Z_i \end{Bmatrix} [dN_i(t) | Y_i, \delta_i, Z_i] - \frac{S^{(1)}_{AIPW}(\boldsymbol{\beta}, T_i)}{S^{(0)}_{AIPW}(\boldsymbol{\beta}, T_i)} E\left[dN_i(t) | Y_i, \delta_{t_i}, Z_i\right]\right]$$

Based on the above expression, it can be seen that the conditional expectation in $A_i(\boldsymbol{\beta}, \boldsymbol{\pi}_i)$ depends on the baseline cumulative hazard and the conditional distribution of $X|T, \delta_t, Z$ The EM algorithm can be used to derive the AIPW estimates.[8] To perform the EM algorithm, the conditional distribution of $X|T, \delta_t, Z$ and the selection probability $\boldsymbol{\pi}$ need to be estimated. It has been shown that if one of them is estimated correctly, the AIPW estimator is consistent (so called double robustness property). Often two parametric working models are used to estimate the conditional distribution of estimate the conditional distribution $X|T, \delta_t, Z$ and the selection probability $\boldsymbol{\pi}$, respectively, and then directly incorporate them into estimation of AIPW estimator. To relax the reliance on the distributional assumptions, nonparametric techniques have been proposed to estimate the conditional distribution and the selection probability. However, as the number of fully observed covariates (i.e. Z) increases, it gets difficult to estimate the conditional distribution and the selection probability nonparametrically. In this paper, we will mainly focus on the performance of the AIPW estimator where two parametric working models are used to estimate the conditional distribution and the selection probability, respectively, and one of the two models is mis-specified. The estimate of standard error for AIPW is derived from 500 bootstrap samples.

3 Nonparametric multiple imputation

Instead of directly incorporating the working models into estimation, we propose to use two working regression models, one for predicting the missing covariates and one for predicting the missing probabilities, to derive two predictive scores to select an imputing set for each missing covariate observation. The two working regression models are only used to derive two predictive scores to select an imputing set. Hence, the approach is expected to be less affected by the mis-specification of the two working models. To conduct nonparametric multiple imputation, for each missing covariate observation we seek an imputing set consisting of subjects who have similar predictive scores as the subject with missing covariate observation. We describe the imputation procedures in detail below.

3.1 Imputation procedures for missing covariate X

3.1.1 Step 1: Estimate the two predictive scores on a Bootstrap sample

To define each imputing set, we first reduce the observed survival data and Z to two scalar indices (predictive scores), which provide an indicator of an individual's value of X and chance of having missing X. White[14] showed that in Cox regression with missing covariates

only under certain cases the conditional distribution of $X|T, \delta_t, Z$ can be exactly specified using cumulative baseline hazard $H_0(t)$, δ_t and Z. Specifically, when both X and Z are binary variables, the conditional distribution of $X|T, \delta_t, Z$ exactly follows a binomial distribution, where $\text{logit}\left[\Pr(X = 1)|T, \delta_t, Z\right] = a_0 + a_1\delta_t + a_2H_0(t) + a_3Z + a_4ZH_0(t)$. When there is no Z, the conditional distribution reduces to $\text{logit}\left[\Pr\left(X = 1|T, \delta_t\right)\right] = a_0 + a_1\delta_t + a_2H_0(t)$. In other cases, only approximate conditional distribution of $X|T, \delta_t, Z$ can be obtained. The approximate conditional distribution depends on cumulative baseline hazard $H_0(t)$, censoring indicator δ_t and the fully observed covariate Z.[14] Hence, all of them will be included in the working regression model for predicting X. To account for potential censoring-ignorable MAR and misspecification of the conditional distribution of $X|T, \delta_t, Z$, we will include the survival outcome data (i.e. Y and δ_t), as well as Z, in the working regression model for predicting the missing probabilities. This strategy summarizes the multi-dimensional structure of the observed survival data and Z into a two-dimensional summary. The hope is that this two-dimensional summary contains most, if not all, the information about the value of missing X and missingness.

Specifically, a linear/generalized linear model with $H_0(t)$, δ_t and Z as the covariates can be fitted to the complete cases to derive a predictive score for X. This score summarizes the relationship between X and $H_0(t)$, δ_t and Z. A logistic regression model with the observed Y, δ_t and Z as the covariates will be fitted to the missing indicator data (i.e. 6_x) to derive a predictive score for missingness. This score summarizes the relationship between missingness and Y, δ_t and Z. The two models will be fitted on a nonparametric bootstrap sample[17] of the original dataset to incorporate the uncertainty of parameter estimates from the working models. This step results in proper multiple imputation (Nielsen[18] and references therein). More specifically, let $\left(Y^B, \delta_t^B, \delta_x^B, Z^B\right)$ denote the bootstrap sample. Two working models are conducted on the bootstrap sample to calculate two predictive scores, $S_x^{(B)}$ and $S_{\delta x}^{(B)}$, for each individual in the bootstrap sample. We further standardize these scores by subtracting their sample mean and dividing by their standard deviation, and denote the standardized scores by $S_x^{c(B)}$ and $S_{\delta x}^{c(B)}$, respectively. Combinations of these two predictive scores will be studied to see to what extent a double robustness property[19] for model mis-specification can be established and whether a robustness property for link function mis-specification can be established for the non-parametric multiple imputation method.

3.1.2 Step 2: Define the imputing set—For subject j with missing X in the original dataset, two predictive scores are derived using the regression coefficient estimates obtained from the bootstrap sample (i.e. $S_x(j)$ and $S_{\delta x}(j)$) and then standardized by subtracting the sample mean of the corresponding bootstrap sample predictive scores and dividing by the standard deviation of the corresponding bootstrap sample predictive scores, respectively (denoted as $S_x^c(j)$ and $S_{\delta x}^c(j)$). The distance between subject j in the original dataset and subject k in the bootstrap sample is then defined as $d(j,k) = \sqrt{w_1\left[S_x^c(j) - S_x^{c(B)}(k)\right]^2 + w_2\left[S_{\delta_x}^c(j) - S_{\delta_x}^{c(B)}(k)\right]^2}$, where w_1 and w_2 are non-negative

weights that sum to one. Non-zero weights for w_2 may be useful in reducing the bias resulting from model mis-specification. Specifically, a small weight w_2 (e.g. 0.2) will result in incorporating the predictive scores from the missing probability model into defining a set of nearest neighbors for subjects with missing X. There are alternative ways to calculate the distance between subjects such as Mahalanobis distance, which accounts for the correlation between the two predictive scores. Once the distance is derived, for subject j, the distance is then employed to define a set of nearest neighbors. This neighborhood consists of NN subjects who have their X observed and have a small distance from subject j in terms of two predictive scores.

3.1.3 Step 3: Impute a value from the imputing set—After the imputing set is defined, a value of X is randomly drawn from the imputing set. Thus, the procedure imputes X only from the subjects with X observed. The non-parametric multiple imputation method based on a nearest neighborhood is denoted as NNMI(NN, w_1, w_2).

3.1.4 Step 4: Repeat Steps 1 to 3 independently M times—Each of the M imputed datasets is based on a different Bootstrap sample. Once the M multiply imputed datasets are obtained, we carry out the MI analysis procedure established in Rubin.[5] Specifically for our purposes, Cox regression analysis with X and Z as the covariates is performed on the M imputed datasets to estimate β_x and β_z. For both β_x and β_z, the final estimate is the average of the M corresponding regression coefficient estimates (i.e. $\hat{\beta}$) and the final variance (denoted $\mathrm{var}\left[\hat{\beta}\right]$) is the sum of the sample variances (denoted as B_β) of the M regression coefficient estimates and the average (denoted as U_β) of the M variance estimates of $\hat{\beta}$. As shown in Rubin,[5] for both β_x and β_z, the quantity $\left[\hat{\beta}-\beta\right]\sqrt{\mathrm{var}\left[\hat{\beta}\right]}$ approximately follows a t distribution with a degree of freedom, $v=(M-1)\left[1+\left[\frac{U_\beta M}{M+1}\right]/B_\beta\right]^2$ We use a value of 10 or higher for M.

4 Illustration of the method on a breast cancer dataset

We demonstrate the nonparametric multiple imputation approach on a dataset which consists of 7050 women diagnosed with stage IV breast cancer between 2005 and 2011 in California. This dataset was extracted from the breast cancer registries under Surveillance, Epidemiology and End Results (SEER) Program. Of the 7050 patients, besides survival data (i.e. survival status and survival time) after diagnosed with breast cancer, for each patient there are several variables collected at diagnosis, as well as Age, Race (Black, White, Other), HER2, Radiation and Surgery. Those variables are summarized in Table 1. HER2 is a member of the human epidermal growth factor receptor family and has been shown to be strongly associated with increased disease recurrence and a poor prognosis for breast cancer patients.[20] According to Table 1, of the 7050 patients, 1293 (18.34%) had missing HER2 value. Table 2 identifies variables predictive of HER2 value and missing probability. Specifically, based on univariate logistic regression analysis for HER2 positive indicator using patients with their HER2 value available (i.e. complete case analysis), Age, Race, Surgery and baseline cumulative hazard, respectively, are predictive of HER2 value and used for performing the PMM method. The results indicate younger patients who did not have

surgery and had a higher hazard rate are more likely to have a positive HER2 value. Based on univariate logistic regression for missing indicator, Age, Surgery, Radiation, survival status (Dead indicator) and baseline cumulative hazard, respectively, are predictive of missing probability. The results indicate older patients who did not have surgery and radiation and had a lower hazard rate are more likely to have a missing HER2 value. Those predictive covariates are then used to derive the conditional distribution of HER2 given the observed data and the selection probability for performing the AIPW estimation and derive two predictive scores for conducting the proposed multiple imputation method. Specifically, a working logistic regression model for HER2 positive indicator with Age, Race and Surgery, as well as survival status and baseline cumulative hazard, as covariates is fitted to derive the conditional distribution of HER2 given the observed data and a HER2 predictive score for each patient. A working logistic regression model for HER2 missing indicator with Age, Radiation and Surgery, as well as survival status and baseline cumulative hazard, as covariates is fitted to derive the selection probability (i.e. π = 1-missing probability) and a predictive score of HER2 missing probability for each patient. To perform the AIPW estimation, the derived conditional distribution of HER2 is then used to derive the conditional expectations and the selection probability is incorporated into the estimation as the weight. To conduct the proposed multiple imputation approach (i.e. NNMI), the two predictive scores are then used to calculate the distance between patients and then select an imputing set for each patient with missing HER2. The number of imputes M is set at 50. Upon the completion of multiple imputation, Cox regression analysis with Age, Black and Others (White as the reference group), HER2, Radiation and Surgery as the covariates is performed on each of the imputed datasets and Rubin's rule[5] is applied to derive the final estimate for each regression coefficient.

The results of the Cox regression estimation for the CC, PMM, AIPW and NNMI methods are provided in Table 3. Table 3 displays the hazard ratio estimate of each covariate along with the associated 95% confidence interval (CI) and *p*-value. The CC and AIPW methods produce similar results. The results indicate that Age, Black and Surgery are significantly associated with survival after diagnosis with stage IV breast cancer. Specifically, older patients tend to have a higher hazard rate than younger patients, Black patients tend to have a higher hazard rate than White patients, and patients without surgery tend to have a higher hazard rate than patients with surgery. Others patients have a slightly lower hazard rate than white patients but not significant at a significance level of 5%. Radiation and HER2 are not significantly associated with survival after diagnosis with stage IV breast cancer. The PMM and NNMI method produces similar results as the CC and AIPW methods, except for Others. The results of PMM and NNMI methods indicate that Others patients have a significantly lower hazard rate than White patients. In addition, the PMM and NNMI methods produce a tighter 95% CI than the CC and AIPW method except for HER2.

5 Simulation study

We perform several simulation studies to investigate the properties of the AIPW, NNMI and PMM methods when Cox regression has a covariate subject to missing and an additional fully observed covariate that is predictive of the missing covariate, and the quantities of interest are the regression coefficients of the Cox regression model. We investigate the

effects of sample size, mis-specification of one of the two working models and mis-specification of the two link functions under a situation with dependent censoring. The simulation program is written in R and is available upon request.

For each of 1000 independent simulated datasets, the predictive covariate Z is generated from a *U(0,1)* distribution. The covariate X subject to missing is generated from either a *Bernoulli[p(Z)]* distribution, where

p(Z) is either based on a logit link (i.e. $p(Z) = \frac{1}{1+e^{\alpha_0+\alpha_1 Z}}$ or a complementary log–log link (i.e. $p(Z) = e^{-e^{\alpha_0+\alpha_1 Z}}$), or a normal distribution with mean a_0+a_1Z and variance s . The failure time T is generated from either an exponential distribution with a hazard rate of $e^{\beta_x X+\beta_z Z}$ or a Weibull distribution with a hazard rate of $\left(e^{\beta_x X+\beta_z Z}\right)\tau t^{\tau-1}$ at time *t*. The censoring time C is also generated from either an exponential distribution with a hazard rate of $e^{\theta_x X+\theta_z Z}$ or a Weibull distribution with a hazard rate of $\left(e^{\theta_x X+\theta_z Z}\right)\gamma t^{\gamma-1}$ at time t. Let Y = min(T,C) and $\delta_t = I(T \le C)$. The missing indicator δ_x ($\delta_x = 1$ if X is observed) if X is observed) is generated from a *Bernoulli[p(Z, Y)]* distribution, where *p(Z, Y)* (i.e. selection probability) is based on a logit link (i.e. $p(Z,Y) = \frac{1}{1+e^{\eta_0+\eta_z Z+\eta_y Y}}$) or a complementary log–log link (i.e. $p(Z,Y) = e^{-e^{\eta_0+\eta z Z+\eta y Y}}$) The regression coefficients and hazard rates are selected to give a desired censoring rate and missing rate.

For the ‘‘Fully-Observed’’ (FO) analysis, treated as the gold standard, we derive Cox regression coefficient estimates for each simulated dataset before any missingness is applied. For the ‘‘Complete-Case’’ (CC) analysis, we derive Cox regression coefficient estimates from the data with X observed. For the AIPW and NNMI methods, a working logistic regression model (denoted by M_1) is fitted to the data with X observed to derive the conditional distribution of X given the observed data and the predictive score of X. A working logistic regression model (denoted by M_2) is fitted to the missing indicator to derive the missing probability and the predictive score of missingness. When both working models include all of the correct covariates in the models (i.e. M_1:Z,δ_t, $\hat{H}_0(t)$; M_2:Z,Y), they are denoted by $AIPW_{11}$ and $NNMI_{11}$, respectively. When the working model for predicting X includes all of the correct covariates but the working model for predicting the missing probability does not (i.e. M_1:Z,δ_t, $\hat{H}_0(t)$; M_2:Z), they are denoted by $AIPW_{12}$ and $NNMI_{12}$, respectively. When the working model for predicting X does not include all of the correct covariates but the working model for predicting the missing probability does (i.e. M_1:Z,δ_t; M_2:Z,Y), they are denoted by $AIPW_{21}$ and $NNMI_{21}$, respectively. When X and δ_x are generated from a complementary log–log model, both AIPW and NNMI methods are considered as mis-specified even if both working models include all of the correct covariates in the models (i.e. $AIPW_{11}$ and $NNMI_{11}$) since the true models are not logit models. The PMM method includes all of the correct covariates for predicting X (i.e. Z , δ_t , $\hat{H}_0(t)$ is

denoted by PMM_1. Based on our prior experience on dealing with missing data (for both missing outcome and missing covariate values) using multiple imputation,[11,21,22] for the NNMI method we set M = 10, NN =5 and (w_1, w_2) = (0.8, 0.2) or (0.2, 0.8).

The results are provided in Tables 4 to 7. The FO analysis, which is the gold standard method, in all situations targets the true values, has the lowest root mean square error (RMSE) and produces coverage rates comparable to the nominal level, 95%. The CC analysis as expected produces biased regression coefficient estimates, especially for the estimate of regression coefficient for Z (i.e. β_z), which results in a much larger RMSE than AIPW and NNMI and a slightly lower coverage rate than the nominal level in some situations due to the bias. When X is binary (Table 4), the PMM_1 method (i.e. all of the correct covariates are included into imputation method) produces reasonable regression coefficient estimates and coverage rates and tends to have a smaller RMSE than AIPW and NNMI in which the missing probability working model is correctly specified. However, when X is continuous (Table 5), the bias of PMM_1 is much larger than AIPW and NNMI in which the missing probability working model is correctly specified.

When both working models include all of the correct covariates (i.e. $AIPW_{11}$ and $NNMI_{11}$), in all situations, both $AIPW_{11}$ and $NNMI_{11}$ methods produce reasonable regression coefficient estimates and coverage rates. The bias of the $NNMI_{11}$ method is comparable to that of the $AIPW_{11}$ method. For both $AIPW_{11}$ and $NNMI_{11}$ methods, the bias for β_x decreases with sample size and the decrease is larger than PMM_1. For $NNMI_{11}$, when X is binary (Table 4) a larger weight on the missing probability predictive score, i.e. (w_1, w_2) = (0.2, 0.8), can reduce the bias of the estimate of b_x but not that of the estimate of β_z. However, when X is continuous (Table 5), a larger weight on the missing probability predictive score increases the bias of both regression coefficient estimates.

When the working logistic regression model for missing indicator d_x (i.e. M_2) is mis-specified (i.e. $AIPW_{12}$ and $NNMI_{12}$), for both binary (Table 4) and continuous (Table 5) X, $NNMI_{12}$ has a smaller bias than $AIPW_{12,}$ especially when X is continuous. For $NNMI_{12}$ with a larger weight on the predictive score for X,i.e. (w_1, w_2) = (0.8, 0.2), the bias decreases with sample size in all situations. When N ¼ 400 and X is binary (Table 4), the bias of $NNMI_{12}$ is comparable to PMM_1. When X is continuous (Table 5), the bias of $NNMI_{12}$ is smaller than PMM_1 in all situations. The bias of $AIPW_{12}$ does not reduce as much as $NNMI_{12}$ with sample size, especially when X is continuous (Table 5). This is because the performance of the AIPW method highly depends on whether a correct model is used to derive the selection probability. Also, in all situations, $AIPW_{12}$'s standard errors tend to underestimate the variability of the regression coefficient estimates, and the underestimate is substantial when X is continuous (Table 5). As a result, $AIPW_{12}$'s coverage rates are lower than the nominal level. $NNMI_{12}$ has a smaller RMSE than $NNMI_{11}$. This is because the mis-specification of missing probability working model induces a much smaller SD for $NNMI_{12}$ even if the bias is larger than $NNMI_{11}$. For $NNMI_{12}$, when X is binary (Table 4) a larger weight on the missing probability predictive score, i.e. (w_1, w_2) = (0.2, 0.8), has a smaller bias than $NNMI_{11}$ when N = 200. However, when sample size is larger or X is continuous (Table 5), a larger weight on the missing probability predictive score increases the bias of both regression coefficient estimates. When the working logistic regression

model for X (i.e. M_1) is mis-specified (i.e. $AIPW_{21}$ and $NNMI_{21}$), for a binary X (Table 4), $NNMI_{21}$ method has a larger bias than $NNMI_{11}$ and $NNMI_{12}$, especially when the sample size is equal to 200. The bias decreases with sample size in all situations. However, for a continuous X (Table 5), in some situations $NNMI_{21}$ method has a smaller bias than $NNMI_{11}$ and $NNMI_{12}$. This is because the working model for predicting a missing continuous covariate is simply based on some approximation. Similar to $NNMI_{11}$, when X is binary (Table 4) for $NNMI_{21}$ a larger weight on the missing probability predictive score, i.e. (w_1, w_2) = (0.2, 0.8), can reduce the bias of the estimate of β_x but not that of the estimate of β_z. However, when X is continuous (Table 5), a larger weight on the missing probability predictive score increases the bias of both regression coefficient estimates. When the link function for both X and δ_x is mis-specified (Table 6), the NNMI methods can still produce reasonable estimates of regression coefficients. This indicates the NNMI method is also robust to mis-specification of the link functions of the two working models. When both T and C are generated from a Weibull distribution (Table 7), the PMM, NNMI and AIPW methods all produce reasonable estimates. This is because they do not need to specify the underlying distributions of failure and censoring times while performing estimation.

In summary, all methods reduce the bias of the standard CC analysis, but the amount of the remaining bias, the efficiency and the validity of the estimated standard errors vary between methods. The performance of the AIPW method depends on whether a correct model is used to derive the selection probability. In contrast, the NNMI method in which two predictive scores are derived from two working regression models can provide reasonable regression coefficient estimates for both X independent and dependent of Z and is robust to mis-specification (the covariates include and the link function) of either one of the two working regression models.

6 Discussion

In this paper we propose a nonparametric multiple imputation approach to handle a missing covariate in Cox regression analysis and compare it with an existing popular AIPW approach. Based on the simulation results, the performance of the AIPW method depends on whether the selection/missing probability model is correctly specified. This indicates while performing the AIPW method, one has to be sure the corresponding model is correct, and specifically requires all aspects of the models including the link functions and choice of covariates to be correct. In contrast, for the nonparametric multiple imputation approach the two working regression modelsare only used to derive two predictive scores to select imputing sets for missing covariate observations. Once the imputing sets are selected, nonparametric multiple imputation procedures are conducted on the sets. Therefore, this approach is expected to have weak reliance on the two working regression models compared to the AIPW method.

The performances of the proposed nonparametric multiple imputation method will depend on the missing rate. Specifically, the missing rate will affect the number of similar ‘‘donors’’ for each missing covariate observation. In a situation with a high missing rate, say, 0.90, a much larger sample size is required for the proposed method to perform well, than a situation with a low missing rate.

As pointed out in the literature,[23–26] when the imputation model is incompatible to the analysis model, multiple imputation may impute covariates that are incompatible with the analysis model and then lead to biased estimates of parameters and the associated variances. To avoid the incompatibility, one can specify a joint model for outcome and covariates for which the conditional distribution of outcome given covariates matches the analysis model and then using the imputation model implied by this joint model.[26] However, it can be challenging to specify the joint model in a situation with missing covariates. The proposed nonparametric multiple imputation does not directly use a statistical model to perform multiple imputation and is, therefore, expected to be less tangible to the incompatibility as long as the right covariates are included in one of the two working models. Based on the numerical results, we do not observe any under-estimation in variation of the parameter estimates and the bias is mainly due to finite sample even if the link function is mis-specified.

In this paper, we assume missingness only depends on the observed data (i.e. MAR mechanism). This assumption is untestable. It is possible that missingness also depends on some unobserved data (i.e. missing not at random mechanism). This indicates non-ignorable missing mechanism may still remain even conditioning on all of the observed data. Sensitivity analysis[27] would be a possible way to evaluate the impact of unobserved data on the proposed multiple imputation approaches. The proposed nonparametric multiple imputation might be less affected by the violation of the MAR assumption since it does not directly use statistical models for performing imputation.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Dr. Chiu-Hsieh Hsu's research was partially supported by the National Cancer Institute grant P30 CA 023074.

References

1. Cox DR. Regression models and life-tables. J Royal Stat Soc Ser B (Methodological) 1972; 34: 187–220.
2. Cox DR. Partial likelihood. Biometrika 1975; 62: 269–276.
3. Andersen PK and Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat 1982; 10: 1100–1120.
4. Little RJA and Rubin DB. Statistical analysis with missing data, 2nd ed. New York, NY: Wiley, 2002.
5. Rubin DB. Multiple imputation for nonresponse in surveys New York, NY: Wiley, 1987.
6. Rathouz PJ. Identifiability assumptions for missing covariate data in failure time regression models. Biostatistics 2007; 8: 345–356. [PubMed: 16840561]
7. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc 1994; 89: 846–866.
8. Wang CY and Chen HY. Augmented inverse probability weighted estimator for Cox missing covariate regression. Biometrics 2001; 57: 414–419. [PubMed: 11414564]
9. Qi L, Wang CY and Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. J Am Stat Assoc 2005; 100: 1250–1263.

10. Kang JDY and Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. Stat Sci 2007; 22: 523–539.
11. Long Q, Hsu C- H and Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. Stat Sinica 2012; 22: 149–172.
12. Rubin DB. Statistical matching using file concatenation with adjusted weights and multiple imputation. J Business Econ Stat 1986; 4: 87–94.
13. Rosenbaum PR and Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985; 39: 33–38.
14. White IR and Royston P. Imputing missing covariate values for the Cox model. Stat Med 2009; 28: 1982–1998. [PubMed: 19452569]
15. Qi L, Wang YF and He Y. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. Stat Med 2010; 29: 2592–2604. [PubMed: 20806403]
16. Marshall A, Altman DG, Royston P, et al. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol 2010; 10: 7. [PubMed: 20085642]
17. Efron B Bootstrap methods: another look at the jackknife. Ann Stat 1979; 7: 1–26.
18. Nielsen SF. Proper and improper multiple imputation. Int Stat Rev 2003; 71: 593607.
19. Robins JM, Rotnitzky A and van der Laan M. Comment on profile likelihood. J Am Stat Assoc 2000; 95: 477–482.
20. Tan M and Yu D. Molecular mechanisms of erbB2-mediated breast cancer chemo-resistance. Adv Experiment Med Biol 2007; 608: 119–129.
21. Hsu C- H, Long Q, Li Y, et al. A nonparametric multiple imputation approach for data with missing covariate values with application to colorectal adenoma data. J Biopharmaceut Stat 2014; 24: 634–648.
22. Hsu C-H, He Y, Li Y, et al. Doubly robust multiple imputation using kernel-based techniques. Biometric J 2016; 58: 588–606.
23. Fay R Proceedings of the section on survey research methods Washington, DC: American Statistical Association, 1992, pp.227–232.
24. Meng XL. Multiple imputation inferences with uncongenial sources of input. Stat Sci 1994; 9: 538–573. (with Discussion).
25. Rubin DB. Multiple imputation after 18 years. J Am Stat Assoc 1996; 91: 473–490.
26. Bartlett JW, Seaman SR, White IR, et al. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Stat Meth Med Res 2015; 24: 462–487.
27. Carpenter JR, Kenward MG and White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Meth Med Res 2007; 16: 259–275.

Table 1.

Data analysis: description of the 7050 stage IV breast cancer patients.

Variable	Mean/Frequency	Standard Deviation/ Percentage
Age	60.91	14.41
Race		
White	5585	79.22
Black	721	10.23
Others	744	10.55
HER2		
Negative	4180	59.29
Positive	1577	23.37
Missing	1293	18.34
Surgery		
No	3916	55.55
Yes	3134	44.45
Radiation		
No	4484	63.60
Yes	2566	36.40

Table 2.

Data analysis: identification of factors associated with missing value and probability of HER2.

Variable	Missing HER2 Value			Missing HER2 Probability		
	OR[a]	95% CI[b]	p[c]	OR	95% CI	p
Age	0.987	(0.983, 0.991)	<0.0001	1.031	(1.026, 1.035)	<0.0001
Black	1.187	(0.983, 1.433)	0.08	1.007	(0.825, 1.229)	0.94
Others	1.360	(1.135, 1.629)	<0.001	0.908	(0.742, 1.112)	0.35
No Radiation	0.913	(0.811, 1.028)	0.13	1.580	(1.385, 1.804)	<0.0001
No Surgery	0.884	(0.787, 0.993)	0.04	2.146	(1.885, 2.443)	<0.0001
Dead	0.997	(0.888, 1.120)	0.96	2.205	(1.937, 2.510)	<0.0001
$H_0(t)$[d]	1.416	(1.235, 1.624)	<0.0001	0.641	(0.549, 0.747)	<0.0001

[a]Odds ratio (HER2+ vs. HER−).

[b]95% Confidence interval.

[c]p-Value.

[d]Baseline cumulative hazard.

Table 3.

Data analysis: results of Cox regression estimation

Variable	**CC**		
	HR[a]	**95% CI**[b]	***p***[c]
Age	1.015	(1.012, 1.017)	<0.01
Black	1.437	(1.286, 1.695)	<0.01
Others	0.887	(0.781, 1.007)	0.06
No Radiation	1.056	(0.978, 1.140)	0.16
No Surgery	1.893	(1.755, 2.042)	<0.01
Her2	0.940	(0.867, 1.020)	0.14

	PMM			AIPW			NNMI(5, 0.8, 0.2)		
Variable	HR	95% CI	*p*	HR	95% CI	*p*	HR	95% CI	*p*
Age	1.018	(1.015, 1.020)	<0.01	1.015	(1.012, 1.017)	<0.01	1.018	(1.015, 1.020)	<0.01
Black	1.443	(1.308, 1.591)	<0.01	1.436	(1.276, 1.615)	<0.01	1.442	(1.307, 1.591)	<0.01
Others	0.879	(0.786, 0.984)	0.03	0.886	(0.776, 1.011)	0.07	0.879	(0.785, 0.983)	0.02
No Radiation	1.044	(0.976, 1.118)	0.21	1.056	(0.981, 1.137)	0.15	1.044	(0.976, 1.118)	0.21
No Surgery	1.885	(1.762, 2.016)	<0.01	1.894	(1.756, 2.042)	<0.01	1.896	(1.773, 2.028)	<0.01
Her2	0.932	(0.860, 1.010)	0.09	0.958	(0.874, 1.049)	0.35	0.939	(0.864, 1.021)	0.14

[a] Hazard ratio.

[b] 95% Confidence interval.

[c] p-Value.

Table 4.

Monte Carlo simulation study: estimation of Cox regression with dependent censoring, where $T \sim \sim \text{Exponential}\left[e^{\text{In}(2)\times\text{-In}(2)Z}\right]$, $C \sim \text{Exponential}\left[e^{-2X+0.IZ}\right]$, $X \sim \text{Bernoulli}\left[p(Z) = \frac{I}{I+e^{0.25-0.5Z}}\right]$ and $\delta_X \sim \text{Bernoulli}\left[p\left(Z,\delta_t,Y\right) = \frac{I}{I+e^{I.5+0.5Z\text{-}2Y}}\right]$.

	β_x = In (2) = 0.693					β_z=−ln(2) = −0.693					
Method	**Est**[a]	**SD**[b]	**SE**[c]	**RMSE**[d]	**CR**[e]	**Est**	**SD**	**SE**	**RMSE**	**CR**	**Div**[f]
N = 200											
FO	0.691	0.191	0.194	0.191	95.3	−0.696	0.333	0.316	0.333	93.7	
CC	0.652	0.358	0.334	0.360	94.1	−0.953	0.563	0.543	0.620	92.4	
PMM_I	0.719	0.337	0.322	0.338	95.4	−0.685	0.342	0.330	0.342	93.9	
$AIPW_{II}$	0.715	0.362	0.319	0.363	91.3	−0.698	0.355	0.408	0.355	97.7	0
$NNMI_{II}$(0.8,0.2)	0.726	0.360	0.343	0.361	94.1	−0.689	0.347	0.333	0.347	94.9	
$NNMI_{II}$(0.2,0.8)	0.705	0.368	0.352	0.368	93.7	−0.674	0.343	0.330	0.344	94.5	
$AIPW_{I2}$	0.735	0.335	0.199	0.338	74.6	−0.765	0.440	0.480	0.446	95.7	187
$NNMI_{I2}$(0.8,0.2)	0.721	0.317	0.313	0.318	96.1	−0.695	0.347	0.332	0.347	94.2	
$NNMI_{12}$(0.2,0.8)	0.716	0.298	0.301	0.299	96.4	−0.697	0.345	0.332	0.345	93.9	
$AIPW_{2I}$	0.714	0.363	0.319	0.364	91.5	−0.698	0.355	0.409	0.355	97.7	0
$NNMI_{2I}$(0.8,0.2)	0.736	0.348	0.339	0.351	94.4	−0.690	0.346	0.332	0.346	94.9	
$NNMI_{2I}$(0.2,0.8)	0.713	0.364	0.351	0.365	93.7	−0.677	0.342	0.330	0.342	94.5	
N = 400											
FO	0.691	0.135	0.136	0.135	95.6	−0.696	0.226	0.220	0.226	94.6	
CC	0.661	0.235	0.23	0.237	94.1	−0.947	0.385	0.370	0.461	89.1	
PMM_I	0.710	0.219	0.225	0.220	95.9	−0.691	0.228	0.230	0.228	95.0	
$AIPW_{II}$	0.701	0.234	0.222	0.234	94.3	−0.698	0.230	0.263	0.230	97.4	0
$NNMI_{II}$(0.8,0.2)	0.703	0.238	0.233	0.238	94.9	−0.692	0.227	0.230	0.227	95.2	
$NNMI_{II}$(0.2,0.8)	0.698	0.241	0.237	0.241	94.4	−0.682	0.225	0.229	0.225	94.9	
$AIPW_{I2}$	0.723	0.224	0.138	0.226	80.1	−0.761	0.281	0.305	0.289	96.1	115
$NNMI_{I2}$(0.8,0.2)	0.709	0.214	0.216	0.215	95.4	−0.699	0.226	0.230	0.226	95.2	
$NNMI_{I2}$(0.2,0.8)	0.719	0.201	0.204	0.203	95.3	−0.704	0.228	0.229	0.228	95.2	
$AIPW_{2I}$,	0.701	0.236	0.222	0.236	93.4	−0.698	0.230	0.263	0.230	97.3	0
$NNMI_{2I}$(0.8,0.2)	0.706	0.235	0.233	0.235	95.3	−0.694	0.227	0.230	0.227	94.9	
$NNMI_{2I}$(0.2,0.8)	0.699	0.241	0.238	0.241	94.7	−0.684	0.226	0.229	0.226	94.9	

Note: Censoring rate: 0.35; Missing rate: 0.63.

[a] Average of 1000 point estimates.

[b] Empirical standard deviation.

[c] Average estimated standard error.

[d] Root mean square error: square root of $\text{bias}^2 + \text{SD}^2$.

[e] Coverage rate of 1000 95% confidence intervals.

[f] Number of disconvergences for AIPW.

Table 5.

Monte Carlo Simulation study: estimation of Cox regression with dependent censoring, where T Exponential $T \sim \sim \text{Exponential}\left[e^{\text{In}(2)\times\text{-In}(2)Z}\right]$, $C \sim \text{Exponential}\left[e^{-2X+0.IZ}\right]$, $X \sim \text{Normal}(0.I5 + Z,I)$ and $\delta_X \sim \text{Bernoulli}\left[p\left(Z,\delta_t,Y\right) = \frac{I}{I + e^{I.5 + 0.5Z\text{-}2Y}}\right]$.

	β_x = In (2) = 0.693					β_Z = −ln(2) = −0.693					
Method	**Est**[a]	**SD**[b]	**SE**[c]	**RMSE**[d]	**CR**[e]	**Est**	**SD**	**SE**	**RMSE**	**CR**	**Div**[f]
N = 200											
FO	0.668	0.1 13	0.1 14	0.1 16	94.4	-0.690	0.318	0.324	0.318	95.8	
CC	0.700	0.168	0.170	0.168	96.1	-0.869	0.446	0.447	0.479	94.4	
PMM_1	0.592	0.123	0.143	0.159	92.5	-0.621	0.329	0.346	0.337	95.2	
$AIPW_{11}$	0.666	0.160	0.151	0.162	93.4	-0.687	0.346	0.373	0.346	96.5	1
$NNMI_{11}$,(0.8,0.2)	0.656	0.150	0.150	0.155	94.3	-0.670	0.337	0.350	0.338	95.6	
$NNMI_{11}$(0.2,0.8)	0.655	0.154	0.153	0.159	94.5	-0.643	0.334	0.349	0.338	95.4	
$AIPW_{12}$	0.902	0.322	0.097	0.384	60.3	-0.814	0.575	0.517	0.588	89.8	1765
$NNMI_{12}$(0.8,0.2)	0.616	0.127	0.142	0.149	92.9	-0.645	0.329	0.347	0.333	95.9	
$NNMI_{12}$(0.2,0.8)	0.594	0.1 16	0.138	0.153	91.8	-0.628	0.324	0.344	0.330	95.4	
$aipw_{21}$	0.667	0.162	0.151	0.164	93.4	-0.688	0.347	0.373	0.347	96.5	1
$NNMI_{21}$(0.8,0.2)	0.656	0.148	0.149	0.153	94.2	-0.667	0.337	0.349	0.338	95.7	
$NNMI_{21}$(0.2,0.8)	0.655	0.154	0.154	0.159	94.1	-0.641	0.334	0.348	0.338	95.4	
N = 400											
FO	0.677	0.080	0.080	0.082	94.1	-0.695	0.224	0.227	0.224	94.9	
CC	0.716	0.120	0.1 18	0.122	95.6	-0.870	0.307	0.309	0.354	91.6	
PMM_1	0.602	0.088	0.099	0.127	85.5	-0.639	0.234	0.243	0.240	95.4	
$AIPW_{11}$	0.676	0.106	0.106	0.107	94.6	-0.695	0.241	0.254	0.241	95.6	0
$NNMI_{11}$ (0.8,0.2)	0.673	0.105	0.106	0.107	94.7	-0.686	0.240	0.244	0.240	95.4	
$NNMI_{11}$(0.2,0.8)	0.673	0.108	0.107	0.110	95.1	-0.667	0.240	0.243	0.241	95.6	
$AIPW_{1\,2}$	0.926	0.224	0.069	0.323	44.0	-0.829	0.371	0.424	0.395	88.7	1596
$NNMI_{12}$(0.8,0.2)	0.631	0.093	0.101	0.112	90.8	-0.663	0.233	0.243	0.235	96.3	
$NNMI1_2$(0.2,0.8)	0.609	0.084	0.098	0.119	87.2	-0.643	0.228	0.242	0.233	95.6	
$aipw_{21}$	0.676	0.107	0.106	0.108	94.6	-0.695	0.241	0.254	0.241	95.6	0
$NNMI_{21}$(0.8,0.2)	0.672	0.105	0.105	0.107	94.9	-0.685	0.240	0.244	0.240	95.7	
$NNMI_{21}$(0.2,0.8)	0.672	0.107	0.107	0.109	94.9	-0.667	0.239	0.243	0.240	95.3	

Note: Censoring rate: 0.33; Missing rate: 0.47.

[a] Average of 1000 point estimates.

[b] Empirical standard deviation.

[c] Average estimated standard error.

[d] Root mean square error: square root of $\text{bias}^2+\text{SD}^2$

[e] Coverage rate of 1000 95% confidence intervals.

[f] Number of disconvergences for AIPW

Table 6.

Monte Carlo simulation study: estimation of Cox regression with dependent censoring, where T ~ Exponential $[e^{\ln(2)X-\ln(2)Z}]$,C ~ Exponential$[e^{-2X+0.1Z}]$, X~Bernoulli$\left[p(Z) = e^{-e^{-1.0+1.5Z}}\right]$ and δ_X ~Bernoulli$\left[p(Z,\delta_t, Y) = e^{-e^{-0.15+0.25Z-Y}}\right]$.

	$\beta_x = \ln(2) = 0.693$					$\beta_z = -\ln(2) = -0.693$					
Method	**Est**[a]	**SD**[b]	**SE**[c]	**RMSE**[d]	**CR**[e]	**Est**	**SD**	**SE**	**RMSE**	**CR**	**Div**[f]
N = 200											
FO	0.682	0.194	0.198	0.194	95.7	−0.710	0.341	0.337	0.341	94.6	
CC	0.662	0.262	0.267	0.264	95.2	−0.846	0.455	0.456	0.480	94.0	
PMM_1	0.686	0.251	0.258	0.251	95.7	−0.721	0.352	0.348	0.353	93.8	
$AIPW_{11}$	0.685	0.245	0.246	0.245	95.0	−0.713	0.354	0.366	0.355	95.6	0
$NNMI_{11}$(0.8,0.2)	0.695	0.250	0.255	0.250	95.7	−0.718	0.352	0.348	0.353	94.2	
$NNMI_{11}$(0.2,0.8)	0.692	0.255	0.256	0.255	94.5	−0.730	0.349	0.346	0.351	94.2	
$AIPW_{12}$	0.700	0.246	0.209	0.246	90.3	−0.726	0.364	0.373	0.365	95.4	74
$NNMI_{12}$(0.8,0.2)	0.694	0.241	0.25	0.241	96.2	−0.712	0.353	0.348	0.354	94.5	
$NNMI_{12}$(0.2,0.8)	0.696	0.232	0.247	0.232	96.8	−0.709	0.351	0.347	0.351	94.0	
$AIPW_{21}$	0.686	0.243	0.246	0.243	94.9	−0.713	0.354	0.366	0.355	95.6	0
$NNMI_{21}$(0.8,0.2)	0.702	0.247	0.252	0.247	95.6	−0.715	0.352	0.347	0.353	94.0	
$NNMI_{21}$(0.2,0.8)	0.696	0.252	0.255	0.252	94.4	−0.730	0.349	0.346	0.351	94.4	
N = 400											
FO	0.688	0.138	0.139	0.138	95.2	−0.701	0.247	0.236	0.247	93.6	
CC	0.670	0.181	0.187	0.182	95.2	−0.832	0.319	0.318	0.348	93.9	
PMM_1	0.688	0.171	0.178	0.171	95.8	−0.707	0.250	0.243	0.250	94.4	
$AIPW_{11}$	0.690	0.171	0.171	0.171	94.9	−0.702	0.251	0.250	0.251	95.2	0
$NNMI_{11}$(0.8,0.2)	0.695	0.178	0.177	0.178	94.4	−0.706	0.252	0.243	0.252	94.2	
$NNMI_{11}$(0.2,0.8)	0.695	0.178	0.178	0.178	94.2	−0.714	0.250	0.242	0.251	94.8	
$AIPW_{12}$	0.701	0.169	0.145	0.169	89.7	−0.717	0.254	0.254	0.255	95.2	65
$NNMI_{12}$(0.8,0.2)	0.696	0.170	0.174	0.170	95.4	−0.699	0.251	0.243	0.251	94.7	
$NNMI_{12}$(0.2,0.8)	0.703	0.164	0.171	0.164	95.8	−0.695	0.250	0.242	0.250	94.8	
$AIPW_{21}$	0.692	0.170	0.171	0.170	95.0	−0.701	0.251	0.250	0.251	95.2	0
$NNMI_{21}$(0.8,0.2)	0.697	0.177	0.176	0.177	94.6	−0.704	0.252	0.243	0.252	94.5	
$NNMI_{21}$(0.2,0.8)	0.697	0.178	0.177	0.178	94.4	−0.688	0.249	0.242	0.249	94.7	

Note: Censoring rate: 0.37; Missing rate: 0.44.

[a] Average of 1000 point estimates.

[b] Empirical standard deviation.

[c] Average estimated standard error.

[d] Root mean square error: square root of $\text{bias}^2 + \text{SD}^2$.

[e] Coverage rate of 1000 95% confidence intervals.

[f] Number of disconvergences for AIPW.

Table 7.

Monte Carlo simulation study: estimation of Cox regression with dependent censoring, where T ~ Weibull($e^{\ln(2)X-\ln(2)Z}$,1.5), C ~ Weibull($e^{-2X+0.1Z}$,1.4), X~Bernoulli$\left[p(Z) = \frac{1}{1+e^{0.25-0.5Z}}\right]$ and δ_X ~Bernoulli$\left[P(Z,\delta_t, Y) = \frac{1}{1+e^{1.5+0.5Z-2Y}}\right]$.

	β_x = ln(2) = 0.693					β_z = ln(2) = 0.693					
Method	**Est**[a]	**SD**[b]	**SE**[c]	**RMSE**[d]	**CR**[e]	**Est**	**SD**	**SE**	**RMSE**	**CR**	**Div**[f]
N = 200											
FO	0.688	0.201	0.195	0.201	94.3	−0.714	0.308	0.316	0.309	95.5	
CC	0.671	0.329	0.320	0.330	94.2	−0.940	0.531	0.518	0.586	92.6	
PMM_1	0.721	0.310	0.303	0.311	96.1	−0.703	0.319	0.330	0.319	95.2	
$AIPW_{11}$	0.705	0.331	0.296	0.331	91.8	−0.716	0.327	0.373	0.328	96.1	0
$NNMI_{11}$(0.8,0.2)	0.717	0.338	0.315	0.339	93.0	−0.705	0.322	0.332	0.322	94.3	
$NNMI_{11}$(0.2,0.8)	0.692	0.339	0.320	0.339	94.1	−0.690	0.317	0.330	0.317	94.3	
$AIPW_{12}$	0.739	0.321	0.208	0.324	81.5	−0.779	0.378	0.450	0.388	96.1	191
$NNMI_{12}$(0.8,0.2)	0.721	0.304	0.296	0.305	95.5	−0.713	0.325	0.331	0.326	94.1	
$NNMI_{12}$(0.2,0.8)	0.720	0.286	0.286	0.287	96.6	−0.714	0.322	0.331	0.323	94.7	
$AIPW_{21}$	0.706	0.331	0.296	0.331	92.1	−0.715	0.327	0.373	0.328	96.2	0
$NNMI_{21}$(0.8,0.2)	0.724	0.331	0.313	0.332	94.0	−0.707	0.322	0.332	0.322	94.1	
$NNMI_{21}$(0.2,0.8)	0.696	0.335	0.319	0.335	94.4	−0.690	0.317	0.329	0.317	94.5	
N = 400											
FO	0.683	0.143	0.136	0.143	93.6	−0.701	0.228	0.221	0.228	95.0	
CC	0.660	0.235	0.220	0.237	92.2	−0.927	0.379	0.358	0.445	88.8	
PMM_1	0.699	0.219	0.217	0.219	95.1	−0.698	0.238	0.231	0.238	94.9	
$AIPW_{11}$	0.691	0.230	0.206	0.230	92.2	−0.706	0.241	0.247	0.241	96.4	0
$NNMI_{11}$(0.8,0.2)	0.697	0.233	0.217	0.233	94.0	−0.700	0.240	0.231	0.240	94.7	
$NNMI_{11}$(0.2,0.8)	0.688	0.234	0.224	0.234	94.2	−0.687	0.236	0.230	0.236	95.2	
$AIPW_{12}$	0.713	0.217	0.144	0.218	81.0	−0.760	0.279	0.270	0.287	94.9	150
$NNMI_{12}$(0.8,0.2)	0.700	0.213	0.205	0.213	94.5	−0.705	0.239	0.230	0.239	95.3	
$NNMI_{12}$(0.2,0.8)	0.708	0.199	0.196	0.200	95.8	−0.708	0.238	0.230	0.238	94.8	
$AIPW_{21}$	0.690	0.230	0.206	0.230	92.3	−0.706	0.240	0.247	0.240	96.3	0
$NNMI_{21}$(0.8,0.2)	0.700	0.232	0.216	0.232	94.1	−0.701	0.240	0.231	0.240	94.6	
$NNMI_{21}$(0.2,0.8)	0.688	0.234	0.221	0.234	94.4	−0.687	0.236	0.229	0.236	95.1	

Note: Censoring rate: 0.35; Missing rate: 0.60.

[a] Average of 1000 point estimates.

[b] Empirical standard deviation.

[c] Average estimated standard error.

[d] Root mean square error: square root of $bias^2 + SD^2$.

[e] Coverage rate of 1000 95% confidence intervals.

[f] Number of disconvergences for AIPW.