Communications for Statistical Applications and Methods
2018, Vol. 25, No. 2, 159–172

https://doi.org/10.29220/CSAM.2018.25.2.159
Print ISSN 2287-7843 / Online ISSN 2383-4757

# Comparison of missing data methods in clustered survival data using Bayesian adaptive B-Spline estimation

Hanna Yoo[a], Jae Won Lee[1,b]

[a]Department of Computer Software, Busan University of Foreign Studies, Korea
[b]Department of Statistics, Korea University, Korea

**Abstract**

In many epidemiological studies, missing values in the outcome arise due to censoring. Such censoring is what makes survival analysis special and differentiated from other analytical methods. There are many methods that deal with censored data in survival analysis. However, few studies have dealt with missing covariates in survival data. Furthermore, studies dealing with missing covariates are rare when data are clustered. In this paper, we conducted a simulation study to compare results of several missing data methods when data had clustered multi-structured type with missing covariates. In this study, we modeled unknown baseline hazard and frailty with Bayesian B-Spline to obtain more smooth and accurate estimates. We also used prior information to achieve more accurate results. We assumed the missing mechanism as MAR. We compared the performance of five different missing data techniques and compared these results through simulation studies. We also presented results from a Multi-Center study of Korean IBD patients with Crohn's disease (Lee *et al.*, *Journal of the Korean Society of Coloproctology*, **28**, 188–194, 2012).

Keywords: Bayesian adaptive B-spline, clustered data, MICE, missing covariates, multiple imputation, single imputation

## 1. Introduction

In epidemiological studies, missing covariate is a common problem in survival data and is often encountered in many statistical applications. The easiest way to deal with missing covariates is to discard missing observations and use only completely observed subjects. However, using only observed subjects may lead to biased estimates of parameters. As the proportion of missing data increases, there will be a substantial loss of efficiency. Rather than discarding missing data, statistical approaches can be applied under the presence of missing covariates. Zhou and Pepe (1995) have proposed an estimated partial likelihood method to estimate relative risk parameters. Lipsitz and Ibrahim (1996) have extended the EM by the method of weights to survival outcomes whose distributions may not fall in the class of generalized linear models. Chen and Little (1999) have used a nonparametric maximum likelihood to estimate regression parameters in a proportional hazards regression model with missing covariates. All these methods mentioned above can be used with the Cox model for univariate survival data with missing covariates. Imputing the missing data can be performed instead of using a likelihood based method to handle missing covariates. Likelihood based methods can be sophisticated. They generally require problem-specific programs. However, imputing the missing data enables the

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, SeongBuk- Gu, Seoul 02841, Korea. E-mail: jael@korea.ac.kr

Published 31 March 2018 / journal homepage: http://csam.or.kr

use of whole data and analysis can be performed easily using existing methods. Marshall *et al.* (2010) have compared several imputing methods (complete case (CC) analysis, single imputation (SI), and five multiple imputation methods) for handling missing covariates in univariate survival data through simulations. With or without imputation, most papers are based on the frequentist maximum likelihood method. Bayesian approaches have also merged and recently applied to survival data more frequently compared to the frequentist maximum likelihood method. Sharef *et al.* (2010) proposed a Bayesian adaptive B-spline estimation when clustered survival data are complete without missing covariates. They modeled the unknown baseline hazard and density of the random effects using a penalized mixture of B-splines. Inclusion of prior information of baseline hazard function and frailty density made the model more flexible. Using the spline component also improved the performance when the hazard or the frailty density was distinctly non-smooth in nature Sharef *et al.* (2010).

In this paper, we extended the Bayesian adaptive B-spline estimation method to clustered survival data with missing covariates. This study compares built-in imputation methods with CC analysis in the presence of missing covariates under clustered survival data using a Bayesian approach. We used the Bayesian adaptive B-spline estimation method proposed by Sharef *et al.* (2010) after imputing the missing covariates. We conducted a simulation study and compared five different missing data techniques: 1) complete case (CC) analysis, 2) single imputation (SI) using predictive mean matching (PMM), 3) MI-MICE, 4) MI-MICE-PMM, and 5) MI-AregImpute (the last three techniques are multiple imputation techniques using multivariate imputation by chained equations (MICE) available within R statistical software). The remainder of the paper is organized as follows. In Section 2, we briefly described each of the five imputation methods. We introduced how we extended the Bayesian adaptive B-spline estimation method when the data had missing covariates in Section 3. Results of the simulation study are summarized in Section 4. We then showed results applying to Crohn's disease data in Section 5. We then concluded the study with a brief discussion in Section 6.

## 2. Missing data methods

Missing covariates are imputed through four different imputation methods with CC analysis. These imputation methods included one SI and three multiple imputation methods. We will briefly describe each method below.

### 2.1. Complete case analysis

The simplest way to handle missing values in the data is to use only observed values (that is, only CC is used). The CC estimator is unbiased when the missing mechanism is missing complete at random (MCAR). However, if there are many variables with missing values, then a large proportion of observations may be dropped. This can cause bias in missing at random (MAR) and lose efficiency in MCAR.

### 2.2. Single imputation using predictive mean matching

Imputing a missing value with a predicted value is called predictive mean matching. It is used to impute a value randomly from a set of observed values whose predicted values are closest to predicted values from a specified regression model (Heitjan and Little 1991; Schenker and Taylor 1996).

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a $n \times p$ matrix composed of $n$ subjects with $p$ variables. Let $Y_i = (Y_{i1}, \ldots, Y_{ip})$ be one of incomplete variables and denote $Y^{\text{obs}} = (Y_1^{\text{obs}}, \ldots, Y_p^{\text{obs}})$ and $Y^{\text{mis}} = (Y_1^{\text{mis}}, \ldots, Y_p^{\text{mis}})$ as observed data and missing data in $Y$, respectively. For each incomplete subject, the con-

ditional expected value $\hat{\mu} = E(Y^{\text{mis}}|Y^{\text{obs}})$ is estimated and the missing value is imputed through the nearest neighborhood subject. When a continuous random variable is imputed, this method is straightforward. However, it might be more complicated for a categorical variable.

## 2.3. Multiple imputation

Multiple imputation is a three-step approach in estimating incomplete data regression models (Rubin, 1987). The notation of $Y$ is the same as described above. $Q$ is denoted as the quantity of interest. The first step is to create plausible values for missing observations from a specifically modeled distribution that can reflect uncertainty about the nonresponse model. Usually, five to ten imputed datasets are created. The imputed dataset is denoted as $Y^{(1)}, \ldots, Y^{(m)}$, where m is the number of imputations. The next step is to analyze the imputed data using typical methods we would have used for complete data. Denote m estimates as $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(m)}$ for each imputed dataset. The last step is to combine results for each imputed dataset and obtain pooled estimate $\bar{Q} = (1/m) \sum_{i=1}^{m} \hat{Q}^{(i)}$. Compared with SI, multiple imputation methods can minimize standard error and increase efficiency of estimates. However, they are more difficult to perform.

The MI-MICE is an imputation method using multivariate imputation by chained equations. MICE is a software for imputing incomplete multivariate data by fully conditional specification. It appeared in the R package library in 2001. Fully conditional method specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each for incomplete variable (van Buuren, 2007). Under the assumption that a multivariate distribution exists from conditional distributions, MICE constructs a Gibbs sampler from specified conditionals. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ denote the vector of $p$ unknown parameters of the multivariate distribution of $Y$. That is, starting from observed marginal distributions, the $i^{th}$ iteration of chained equations is a Gibbs sampler that successively draws.

$$\begin{aligned}
\theta_1^{*(t)} &\sim P\left(\theta_1|Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}\right), & Y_1^{*(t)} &\sim P\left(Y_1|Y_1^{\text{obs}}, Y_2^{(t-1)}, \ldots, Y_p^{(t-1)}, \theta_1^{*(t)}\right), \\
&\vdots & &\vdots \\
\theta_p^{*(t)} &\sim P\left(\theta_p|Y_p^{\text{obs}}, Y_1^{(t-1)}, \ldots, Y_{p-1}^{(t-1)}\right), & Y_p^{*(t)} &\sim P\left(Y_p|Y_p^{\text{obs}}, Y_1^{(t-1)}, \ldots, Y_{p-1}^{(t-1)}, \theta_p^{*(t)}\right),
\end{aligned}$$

where $Y_i^{(t)} = (Y_i^{\text{obs}}, Y_i^{(t)})$ is the $i^{th}$ imputed variable at iteration $t$. This method has different names, including regression switching (van Buuren *et al.*, 1999), variable-by-variable imputation (Brand, 1999), and chained equations. In this paper, we used imputation methods built in the MICE software in the R package library.

MI-MICE-PMM is a method that imputes missing values under multiple imputation with regression switching (MICE) using predictive mean matching (PMM). The imputing method is the same as the SI except that the PMM imputation method is done in $m$ imputed datasets $Y^{(1)}, \ldots, Y^{(m)}$.

Lastly, the flexible additive imputation model with PMM under multiple imputation method (MI-AregImpute) is performed under the library 'aregImp' in the MICE package. This method takes all aspects of uncertainty in imputations into account by using bootstrap to approximate the process of drawing predicted values from a full Bayesian predictive distribution (Frank, 2010). For each multiple imputation, different bootstrap resamples are used and a flexible additive model is fitted on a sample with replacement from the original data. This model is then used to predict all original values for the target variable.

## 3. Statistical model

Clustered failure time data occur when study subjects from the same cluster share common characteristics such that failure times within the same cluster are correlated. One possible mechanism is the existence of a common risk. When this common risk acts as a factor on hazard function, it is called frailty. The frailty model was first named by Vaupel *et al.* (1979). It was studied by Clayton (1978) for multivariate failure time data. It is usually used for clustered failure time data.

Let $X_{ij}$ and $C_{ij}$ denote the failure and censoring time of the $j(=1,\ldots,m_i)^{th}$ subject in the $i(=1,\ldots,M)^{th}$ cluster, respectively. Furthermore, let $Z_{ij}$ be a $p$ vector fixed covariates. We first assumed that $Z_{ij}$ had no missing values. The observed subject took the form $T_{ij} = \min(X_{ij}, C_{ij})$ and $\delta_{ij} = I(X_{ij} \leq C_{ij})$.The correlation between subjects in the same cluster was incorporated through random effects or frailties $U_i (i = 1,\ldots,M)$. Frailties represent the heterogeneity of each cluster. They are assumed to have a particular distribution $f(\cdot)$. We assumed that censoring time was conditionally independent of the failure time given covariates and frailties. The hazard function for subject $j$ in cluster $i$ is given by the following:

$$\lambda_{ij}(t|\boldsymbol{U},\boldsymbol{Z}) = U_i\lambda_0(t)\exp\left(Z_{ij}^t\boldsymbol{\beta}\right), \tag{3.1}$$

where $\boldsymbol{U} = (U_i,\ldots,U_M)^T$, $\boldsymbol{Z} = \{Z_{ij}, i,\ldots,M; j = 1,\ldots,m_i\}$, and $\boldsymbol{\beta}$ is a $p$-dimensional vector of regression coefficients.

In this study, we used mixtures of B-spline for the baseline hazard $\lambda_0(\cdot)$ and the density of the frailty $f(\cdot)$ in formula (3.1) as noted in Sharef *et al.* (2010). We followed their notations. Markov Chain Monte Carlo (MCMC) was then used to obtain regression coefficients. For the baseline hazard function $\lambda_0(\cdot)$, a non-negative linear combination of $K_\lambda$ B-spline basis functions $B_{\lambda k}(\cdot)$ was used. For $t \geq 0$, the baseline hazard function was presented as

$$\lambda_0\left(t|\boldsymbol{\theta_\lambda},\boldsymbol{\eta_\lambda},\phi_\lambda\right) = \lambda_{0p}(t|\boldsymbol{\eta_\lambda}) + \phi_\lambda\left[\sum_{k=1}^{K_\lambda} B_{\lambda k}(t)\omega_{\lambda k} - \lambda_{0p}(t|\boldsymbol{\eta_\lambda})\right], \tag{3.2}$$

where $\boldsymbol{\theta_\lambda}$ is the spline parameter with weight $\omega_{\lambda k} = e^{\theta\lambda k}$ for each basis function, $\boldsymbol{\lambda_{0p}}(\cdot|\boldsymbol{\eta_\lambda})$ denotes the hazard function for a specified parametric family, and $\boldsymbol{\eta_\lambda}$ is the parameter of the parametric family. Including a parametric part enabled us to model the baseline hazard and frailty density curve smoother with less variable fits. Weibull or lognormal distribution is widely used. $\phi_\lambda \in [0,1]$ is the weight that specifies the degree of confidence in the parametric component. For example, $\phi_\lambda = 0$ will lead to a purely parametric Bayesian survival model.

Next, the frailty density function can be written as

$$f(x|\boldsymbol{\theta_u},\boldsymbol{\eta_u},\phi_u) = \phi_u\left[\sum_{k=1}^{K_u}\tilde{B}_{uk}(x)\omega_{uk} + (1-\phi_u)f_p(x|\boldsymbol{\eta_u})\right], \tag{3.3}$$

where $\boldsymbol{\theta_u}$ is the spline parameter and $f_p(\cdot|\boldsymbol{\eta_u})$ is the density function for a specified parametric family of probability distributions with parameters $\boldsymbol{\eta_u}$. The normalized B-spline basis function is written as $\tilde{B}_{uk}(x) = B_{uk}(x)\cdot\left(\int_{-\infty}^{\infty} B_{uk}(s)ds\right)^{-1}$ for $x \geq 0$ with weight $\omega_{uk} = \{\exp(\theta_{uk})/\sum_{l=1}^{Ku}\exp(\theta_{ul})\}\cdot\phi_u \in [0,1]$ is the weight for the parametric component.

For the regression parameter $\boldsymbol{\beta}$ and spline parameters $\boldsymbol{\theta_\lambda}$ and $\boldsymbol{\theta_u}$, multivariate Gaussian distribution with variances $\sigma_\beta^2, \sigma_\lambda^2, \sigma_u^2$ is assumed, respectively. They are independent of each other. The prior distribution of $\sigma_\beta^2, \sigma_\lambda^2, \sigma_u^2$ is assumed to be multivariate Gaussian. Penalty functions $p_\lambda(\boldsymbol{\theta_\lambda})$ and $p_u(\boldsymbol{\theta_u})$

are incorporated to induce smoothness in B-spline coefficients and avoid over fitting. For weights $\boldsymbol{\phi} = (\phi_\lambda, \phi_u)$ in formula (3.2) and (3.3), Beta priors with hyperparameters $\alpha_{\phi_u}$ and $\alpha_{\phi_\lambda}$ can be used.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta_\lambda}, \boldsymbol{\theta_u})$ denote the set of spline parameters and $\boldsymbol{\sigma} = (\sigma_\beta^2, \sigma_\lambda^2, \sigma_u^2)$ denote the set of variance parameters. For parametric distribution, denote the set of parameters as $\boldsymbol{\eta} = (\boldsymbol{\eta_\lambda}, \boldsymbol{\eta_u})$ with prior distribution $\pi_\lambda(\eta_\lambda|\tau_\lambda)$ and $\pi_u(\eta_u|\tau_u)$, respectively. The prior set is denoted as $\boldsymbol{\tau} = (\boldsymbol{\tau_\lambda}, \boldsymbol{\tau_u})$. It may have hyperparameters $\boldsymbol{\alpha_\lambda} = (\alpha_{\eta\lambda}, \alpha_{\eta u})$. The log of the posterior density is then given by

$$
\begin{aligned}
& l(\boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\tau}|\boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{Z}) \\
& = \sum_{i,j} \delta_{ij} \left[ \log(U_i) + \log \lambda_0(T_{ij}|\boldsymbol{\theta_\lambda}, \boldsymbol{\eta_\lambda}, \phi_\lambda) \right] \\
& \quad - \sum_{i,j} \delta_{ij} U_i \Lambda_0(T_{ij}|\boldsymbol{\theta_\lambda}, \boldsymbol{\eta_\lambda}, \phi_\lambda)^{e^{\boldsymbol{z_{ij}^T \beta}}} + \sum_i \log f(U_i|\boldsymbol{\theta_u}, \boldsymbol{\eta_u}, \phi_u) \\
& \quad - \frac{p}{2} \log \sigma_{\beta^2} + \frac{\boldsymbol{\beta^T \beta}}{2\sigma_{\beta^2}} - \sum_{d \in \{\lambda, u\}} \left\{ \frac{K_d}{2} \log \sigma_{d^2} + \frac{p_d(\boldsymbol{\theta_d})}{2\sigma_{d^2}} \right\} - \sum_{d \in \{\beta, \lambda, u\}} \left\{ (\alpha_{d1} + 1) \log \sigma_{d^2} + \frac{\alpha_{d^2}}{\sigma_{d^2}} \right\} \qquad (3.4) \\
& \quad + \log \left\{ \phi_\lambda^{(\alpha_{\phi_{\lambda 1}} - 1)} (1 - \phi_\lambda)^{(\alpha_{\phi_{\lambda 2}} - 1)} \right\} \\
& \quad + \log \left\{ \phi_u^{(\alpha_{\phi_{u1}} - 1)} (1 - \phi_u)^{(\alpha_{\phi_{u2}} - 1)} \right\} \qquad (3.5) \\
& \quad + \log \pi_\lambda \left( \boldsymbol{\eta_\lambda} | \boldsymbol{\tau_{\eta_\lambda}} \right) + \log \pi_u \left( \boldsymbol{\eta_u} | \boldsymbol{\tau_{\eta_u}} \right) \\
& \quad + \log \pi_\eta \left( \boldsymbol{\eta_u} | \boldsymbol{\tau_{\eta_u}} \right), \qquad (3.6)
\end{aligned}
$$

where (3.4) corresponds to log likelihood $l(\boldsymbol{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{T}, \boldsymbol{\delta}, \boldsymbol{Z})$, (3.5) corresponds to Beta priors on weights, and the term (3.6) corresponds to the priors on parametric components. By maximizing the conditional log-likelihood, spline and parametric coefficients $\boldsymbol{\theta_\lambda}, \boldsymbol{\theta_u}, \boldsymbol{\eta_\lambda}, \boldsymbol{\eta_u}$ are initialized. After initial values are obtained, each set of parameters is updated by Gibbs sampling and Metropolis_Hastings MCMC.

When there is no missing covariate, parameters of interest can be obtained as shown above. Since there are missing covariates in the data, we used imputation methods to fill in the missing values. In addition, we also performed CC analysis and obtained coefficients of parameters. Let the complete data be noted by $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where $Y_{\text{obs}}$ and $Y_{\text{mis}}$ are observed part and missing part of $Y$, respectively. Also, let $P(Y|\theta)$ model be the complete data and $\theta$ denote parameters of interest. The posterior predictive distribution for $Y_{\text{mis}}$ is given by:

$$
P(Y_{\text{mis}}|Y_{\text{obs}}) = \int P(Y_{\text{mis}}|Y_{\text{obs}}, \theta) P(\theta|Y_{\text{obs}}) d\theta, \qquad (3.7)
$$

where

$$
P(\theta|Y_{\text{obs}}) \propto P(\theta) \int P(Y_{\text{obs}}, Y_{\text{mis}}|\theta) dY_{\text{mis}}. \qquad (3.8)
$$

The basic scheme of the process is given as:

Step 1: For each cluster $i = 1, \ldots, M$, repeat the following steps for $j = 1, \ldots, Q$ plausible imputed data sets.

1. Imputation step: Generate missing values $Y_{\text{mis}}^{j+1}$ from $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta^j)$.

2. Posterior step: Draw parameters $\theta^{j+1}$ from $P(\theta|Y_{\text{obs}}, Y_{\text{mis}}^{j+1})$.

Repeating the above steps will generate the following Markov Chain

$$\left(Y_{\text{mis}}^1, \theta^1\right), \left(Y_{\text{mis}}^2, \theta^2\right), \ldots, \left(Y_{\text{mis}}^j, \theta^j\right). \tag{3.9}$$

These two steps are iterated with a starting value $\theta^0$ until the distribution $P(Y_{\text{mis}}|Y_{\text{obs}}, \theta)$ is stabilized.

Step 2: For $Q$ imputed data sets, model the baseline hazard function and the frailty density function as formula (3.2) and (3.3), respectively. By applying Bayesian method, estimate coefficients estimator of interest in each $Q$ imputed data sets and denote them as $\hat{\theta}^{(\mathbf{1})}, \hat{\theta}^{(\mathbf{2})}, \ldots, \hat{\theta}^{(\boldsymbol{Q})}$.

Step 3: Combine $Q$ estimates $\hat{\theta}^{(\mathbf{1})}, \hat{\theta}^{(\mathbf{2})}, \ldots, \hat{\theta}^{(\boldsymbol{Q})}$ into $\hat{\theta}$ which is the mean of $Q$ estimates. For CC analysis, subjects with missing values are deleted. For SI, only step 1 is needed.

## 4. Simulation studies

We conducted a simulation study to investigate the performance of various missing data methods under several different data settings. Three covariates $Z_{ij1}, Z_{ij2}, Z_{ij3}$ ($i = 1, \ldots, M$, $j = 1, \ldots, m_i$) are considered, where $Z_{ij1}$ has no missing values and $Z_{ij2}, Z_{ij3}$ have missing values. We assumed the missing mechanism as MAR. Therefore, the missingness in covariates $Z_{ij2}, Z_{ij3}$ only depends on observed values. We set the distribution of $Z_{ij1}$ and $Z_{ij3}$ as $\binom{Z_1}{Z_3} \sim \text{MVN}\left(\binom{2}{1}, \binom{0.50\ \ 0.15}{0.15\ \ 0.25}\right)$ and assumed $Z_{ij2}$ had a Bernoulli distribution with mean $p = \exp(-1 + 2 \times Z_{ij1})/\{1 + \exp(-1 + 2 \times Z_{ij1})\}$. Frailties $U_i$ were sampled from $LN(1, 0.25^2)$. We set coefficients as $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3) = (1, -1, 1)$. We fixed hyperparameters $\alpha = 0.01$ and all variances of frailties as $\sigma_\lambda^2 = \sigma_u^2 = 0.1$ to achieve suitably noninformative priors. For each cluster $i$, sample $m_i$ clustered failure times $X_{ij}$ from exponential distribution with intensity $U_i \exp(\beta_1 Z_{ij1} + \beta_2 Z_{ij2} + \beta_3 Z_{ij3})$. Censoring time $C_{ij}$ was also generated from exponential distribution with appropriate mean which resulted in 20% of censoring rate. Then the observed failure time was obtained as $T_{ij} = \min(X_{ij}, C_{ij})$ and the censoring indicator as $\delta_{ij} = I(X_{ij} \leq C_{ij})$. We set the number of clusters as $M = 10$ and 100 with cluster size as $m_i = 50$ and 5, respectively. Therefore, the whole sample size was 500. For Weibull baseline and lognormal frailty distribution, we set a Beta(1, 2) prior on the weight.

Missing data were then generated as:

$$p(r_{ijk} = 1|\boldsymbol{x_{ij}}, \boldsymbol{z_{ijk}}, \rho) = \frac{\exp(\rho_0 + \rho_1 h(x_{ij}) + \rho_2 z_{ij1} \times h(x_{ij}))}{1 + \exp(\rho_0 + \rho_1 h(x_{ij}) + \rho_2 z_{ij1} \times h(x_{ij}))}, \quad k = 2, 3, \tag{4.1}$$

where $r_{ijk}$ indicates whether $z_{ijk}$ is missing and $h(x_{ij})$ is an indicator function which is 1 if $x_{ij}$ is in the last quartile of the survival time and 0 otherwise. $\boldsymbol{\rho} = (\rho_0, \rho_1, \rho_2) = (-5.45, -2.50, 3)$, $(-4.71, -2.14, 3)$, $(-3.24, -1.39, 3)$, and $(-1.46, -0.68, 3)$, for 5%, 10%, 25%, and 50% missing, respectively. We replicated the simulation 300 times. Table 1 shows the bias of the regression coefficient estimate $\hat{\boldsymbol{\beta}}$ and the estimated standard deviation for frailty $\hat{\sigma}_u$ under various missing rates with five different missing data methods when $M = 10$ and $m_i = 50$. Figure 1 shows a summary of Table 1. We denote the notation of covariate $Z_{ij1}, Z_{ij2}, Z_{ij3}$ as $Z1, Z2$, and $Z3$ respectively. For covariate $Z1$, the missing rate corresponds to the whole data missing rate since $Z1$ has no missing value. We wanted to see how the bias of $\hat{\beta}_1$ was affected by the missing rate of the whole data. CC has the smallest bias among all other imputation methods when a missing rate is rather small. However, when the missing rate is greater than 10%, the

Table 1: Bias of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u$ under various missing rates using five different missing data methods when $M = 10$ and $m_i = 50$

| | Missing rate | Bias of $\hat{\boldsymbol{\beta}}$ | | | Bias of $\hat{\sigma}_u$ |
|---|---|---|---|---|---|
| | | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | |
| CC | 5% | 0.011 | −0.051 | −0.052 | −0.005 |
| | 10% | 0.030 | 0.022 | −0.104 | −0.012 |
| | 25% | −0.070 | 0.174 | −0.239 | −0.024 |
| | 50% | −0.377 | 0.299 | −0.348 | −0.034 |
| SI | 5% | −0.114 | −0.024 | −0.053 | 0.006 |
| | 10% | −0.130 | 0.053 | −0.108 | 0.001 |
| | 25% | −0.128 | 0.157 | −0.217 | 0.012 |
| | 50% | −0.138 | 0.352 | −0.266 | −0.019 |
| MICE | 5% | −0.114 | −0.025 | −0.051 | 0.006 |
| | 10% | −0.133 | 0.054 | −0.107 | 0.003 |
| | 25% | −0.131 | 0.156 | −0.216 | 0.006 |
| | 50% | −0.143 | 0.336 | −0.255 | −0.017 |
| MICE-PMM | 5% | −0.115 | −0.031 | −0.051 | 0.006 |
| | 10% | −0.134 | 0.045 | −0.106 | 0.002 |
| | 25% | −0.128 | 0.163 | −0.215 | 0.007 |
| | 50% | −0.143 | 0.322 | −0.258 | −0.019 |
| AregImpute | 5% | −0.037 | −0.049 | −0.051 | 0.000 |
| | 10% | −0.015 | 0.003 | −0.098 | −0.009 |
| | 25% | 0.006 | 0.096 | −0.230 | −0.015 |
| | 50% | −0.146 | 0.176 | −0.351 | −0.036 |

bias is increased rapidly. This shows the explicit drawback of CC analysis. AregImpute showed the best performance for a missing rate of 10% and 15%. SI, MICE, and MICE-PMM imputation methods showed rather unbiased results when missing rate was less than 50%. Especially for AregImpute, the bias decreased even when the missing rate increased up to 25%. This method had the smallest bias among the five methods for almost every missing rate tested.

For discrete covariate $Z2$, at missing rate up to 5%, the SI method had the smallest bias while CC had the largest bias. This conflicted with the result of covariate $Z1$, where CC had the smallest bias under a 5% missing rate. SI, MICE, and MICE-PMM showed similar results. However, SI had the largest bias when the missing rate was 50%. AregImpute had the smallest overall bias among all missing data methods. For continuous variable $Z3$, with missing rate up to 25%, all methods showed similar results with fine estimate results. However, when the missing rate was increased, CC and AregImpute had the largest bias while MICE and MICE-PMM methods showed the best performance.

Table 2 shows the bias of regression coefficient estimates $\beta_1, \beta_2, \beta_3$ and the estimated standard deviation for frailty $\hat{\sigma}_u$ when the number of cluster is $M = 100$ and the cluster size is $m_i = 5$. Figure 2 summarizes the data in Table 2 as a graph. For all three covariates, performances of the five different missing data methods shown in Figure 2 are similar to those in Figure 1. For covariate $Z1$, the AregImpute method showed the smallest bias among all missing data methods.

For covariate $Z2$, MICE, and MICE-PMM showed a good performance at missing rate below 10%. However, the AregImpute method showed the best performance when the missing rates are above 25%. For covariate $Z3$, the SI method had the smallest bias for all missing rates. This shows that the SI method is not inferior to MI methods. However, CC showed the worst performance among all missing data methods at all missing rates.

In Figure 3, the bias of the frailty standard deviation estimate $\hat{\sigma}_u$ is shown under the same settings of Figure 1 and Figure 2. Figure 3(a) shows regression coefficient estimates of the frailty when the number of clusters and cluster size are 10 and 50, respectively. Figure 3(b) shows regression

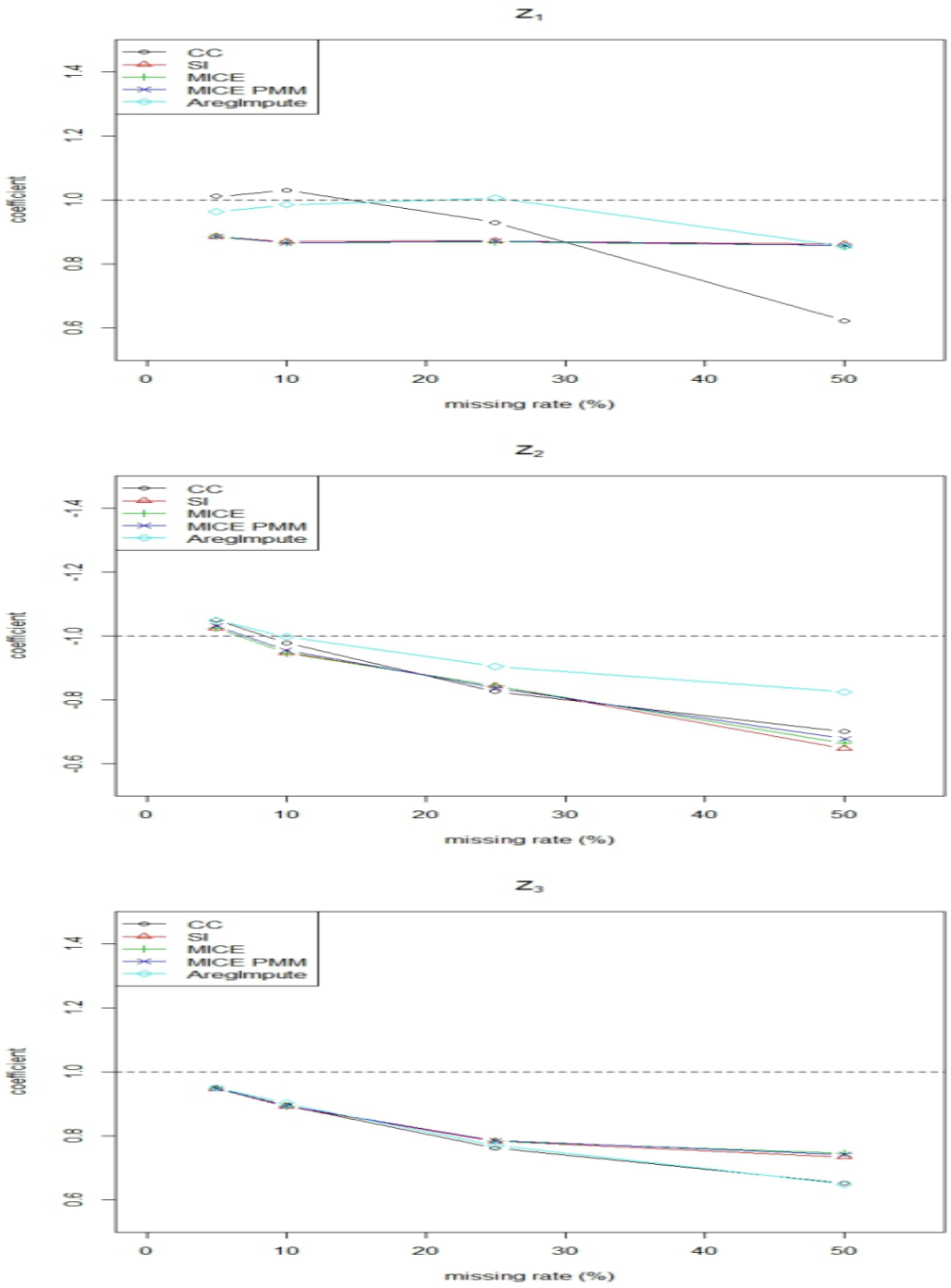



Figure 1: *Regression coefficient estimates using five different missing data methods when* $M = 10$ *and* $m_i = 50$ *under various missing rates.*

Table 2: Bias of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_u$ under various missing rates using five different missing data methods when $M = 100$ and $m_i = 5$

| | Missing rate | Bias of $\hat{\boldsymbol{\beta}}$ | | | Bias of $\hat{\sigma}_u$ |
|---|---|---|---|---|---|
| | | $\hat{\beta}_1$ | $\hat{\beta}_2$ | $\hat{\beta}_3$ | |
| CC | 5% | 0.025 | −0.077 | −0.043 | 0.113 |
| | 10% | 0.071 | −0.002 | −0.086 | 0.121 |
| | 25% | −0.026 | 0.124 | −0.217 | 0.151 |
| | 50% | −0.338 | 0.223 | −0.299 | 0.214 |
| SI | 5% | −0.092 | −0.047 | −0.026 | 0.118 |
| | 10% | −0.087 | 0.010 | −0.064 | 0.131 |
| | 25% | −0.087 | 0.108 | −0.165 | 0.139 |
| | 50% | −0.103 | 0.297 | −0.202 | 0.122 |
| MICE | 5% | −0.095 | −0.044 | −0.029 | 0.120 |
| | 10% | −0.085 | 0.008 | −0.070 | 0.130 |
| | 25% | −0.090 | 0.098 | −0.173 | 0.132 |
| | 50% | −0.106 | 0.296 | −0.212 | 0.120 |
| MICE-PMM | 5% | −0.098 | −0.037 | −0.030 | 0.120 |
| | 10% | −0.094 | 0.020 | −0.068 | 0.126 |
| | 25% | −0.095 | 0.098 | −0.172 | 0.134 |
| | 50% | −0.110 | 0.303 | −0.205 | 0.118 |
| AregImpute | 5% | −0.021 | −0.078 | −0.036 | 0.112 |
| | 10% | 0.034 | −0.040 | −0.078 | 0.125 |
| | 25% | 0.037 | 0.047 | −0.200 | 0.140 |
| | 50% | −0.114 | 0.143 | −0.298 | 0.182 |

coefficient estimates of the frailty when the number of clusters and the cluster size are 100 and 5, respectively. We can see that the bias is affected by the cluster size rather than the number of clusters since Figure 3(a) shows a rather small bias, with SI and MICE-PMM showing the smallest bias. However, the bias is rather large in Figure 3(b).

## 5. Example

We applied the missing data methods to analyze the data from biopsy-proven Crohn's Disease patients who underwent abdominal surgery from January 2000 to December 2009 (Lee *et al.*, 2012). Crohn's disease (CD) is heterogeneous in nature. The only consistent factor is its inconsistency. There have been studies revealing risk factors for post-operative recurrences. However, only a few studies have evaluated risk factors for reoperation after the primary surgery in CD patients. The primary outcome of the study was time to have re-operation from CD recurrence. Data were collected from 627 patients in 18 different hospitals. Each hospital corresponded to a cluster. Patients in each cluster corresponded to cluster subjects. The number of cluster size ranged from 2 to 150 and the median size was 15. The following covariates were considered as risk factors: age at diagnosis, tumor location, and tumor behavior. These three variables were considered as Montreal classification variables. All these variables were categorical variables. We provided the basic description of each covariate and its missing rate in Table 3. We used four different imputation methods with CC method and obtained regression coefficients using a Bayesian adaptive B-spline method. For baseline hazard, a Weibull model was incorporated in the parametric model part. For the frailty density model, lognormal distribution was assumed. For model selection criterion, we used a modification version of the Deviance Information Criterion (Spiegelhalter *et al.*, 2002). DIC is computed as

$$\text{DIC} = \bar{D} + \frac{V}{2}, \tag{5.1}$$

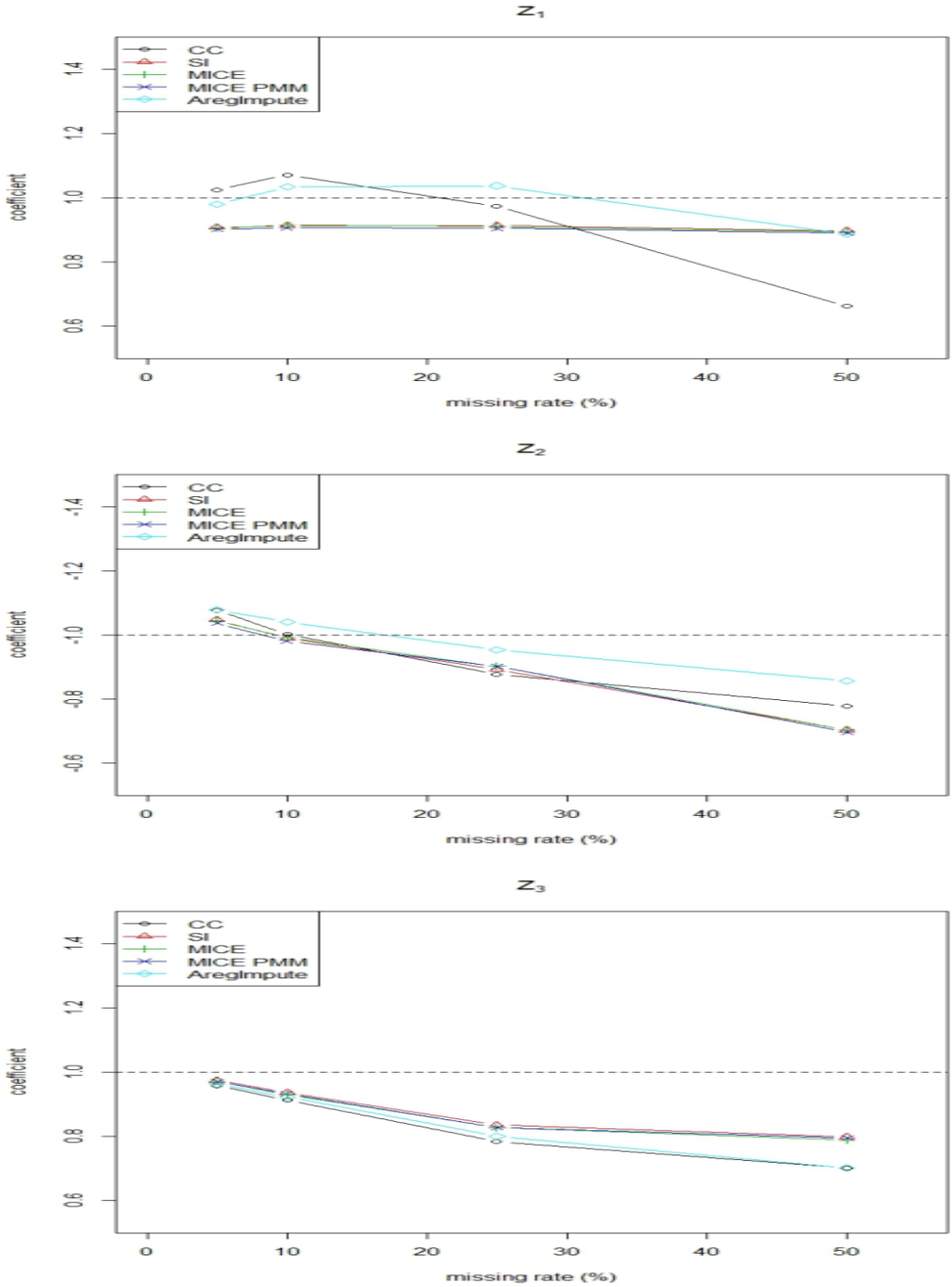
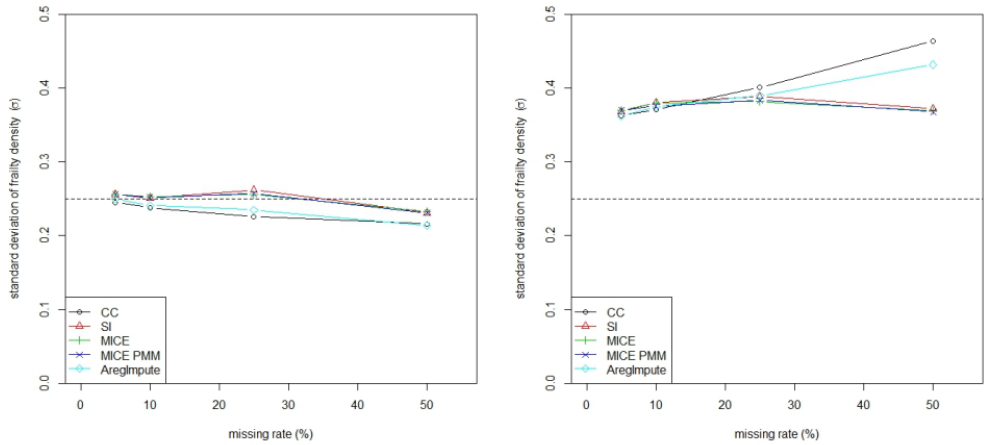



Figure 2: *Regression coefficient estimates using five different missing data methods when* $M = 100$ *and* $m_i = 5$ *under various missing rates.*




(a) $M = 10$ and $m_i = 50$ (b) $M = 100$ and $m_i = 5$

Figure 3: *Regression coefficient estimates of the frailty under five different missing data methods with various missing rates.*

Table 3: Covariates and basic description and missing rates for the CD data

| Name | Description | Missing rate |
|---|---|---|
| Age at diagnosis | A1 less than 16 years<br>A2 between 17 and 40 years<br>A3 older than 40 years | 0.3% |
| Tumor location | L1 ileal<br>L2 colonic<br>L3 ileocolonic<br>L4 isolated upper disease | 8.1% |
| Tumor behavior | B1 non-stricturing, non-penetrating<br>B2 stricturing<br>B3 penetrating | 8.1% |

Table 4: Deviance Information Criterion estimates using five different missing data methods

| CC | SI | MICE | MICE-PMM | AregImp |
|---|---|---|---|---|
| 1841.091 | 1637.391 | 1757.315 | 1716.785 | 1809.242 |

where $\bar{D}$ and $V$ are the sample mean and the variance of deviance calculated at each MCMC step, respectively. We run the sampling chain for 1500 iterations, discarding the first 500 iterates as burn-in and thinning the chain to every $10^{th}$ sample.

In Table 4, DIC showed that SI (DIC = 1637.391) or MICE-PMM (DIC = 1716.785) method was a good choice for missing value imputation. These results are consistent with the simulation study. The real data coincides with the simulation case where $M = 10$, $m_i = 50$ at missing rates of 5% and 10%. In this case, for categorical variable $Z2$, CC had a larger bias than the other four methods while SI had the smallest bias.

Table 5 shows the posterior mean and 95% posterior intervals for covariate effects and the standard deviation of the frailty. In all five imputation methods, signs and magnitudes of regression coefficients

Table 5: Posterior mean and 95% credible intervals of regression coefficients and frailty variance for the CD data using five missing data methods

| Covariates | CC | | | MICE-PMM | | |
|---|---|---|---|---|---|---|
| | Posterior mean | 2.5% quantile | 97.5% quantile | Posterior mean | 2.5% quantile | 97.5% quantile |
| Age_linear | −0.1366 | −0.5951 | 0.1619 | −0.1210 | −0.5047 | 0.2270 |
| Age_quad | −0.0455 | −0.3359 | 0.2131 | −0.0881 | −0.3957 | 0.2028 |
| Location_L2 | 0.2474 | −0.1645 | 0.7326 | 0.2067 | −0.1260 | 0.6014 |
| Location_L3 | 0.0931 | −0.2501 | 0.4536 | 0.0425 | −0.2486 | 0.3089 |
| Location_L4 | 0.0833 | −0.2730 | 0.5939 | 0.1457 | −0.2504 | 0.6150 |
| Behavior_B2 | −0.0897 | −0.4712 | 0.2559 | −0.0960 | −0.4437 | 0.1797 |
| Behavior_B3 | −0.0765 | −0.4155 | 0.3961 | −0.0954 | −0.4696 | 0.2143 |
| Variance of frailties | 0.5236 | 0.1492 | 1.2387 | 0.5227 | 0.1609 | 1.2889 |
| Covariates | SI | | | AregImp | | |
| | Posterior mean | 2.5% quantile | 97.5% quantile | Posterior mean | 2.5% quantile | 97.5% quantile |
| Age_linear | −0.1595 | −0.7160 | 0.2415 | −0.1070 | −0.4810 | 0.2161 |
| Age_quad | −0.1035 | −0.5017 | 0.2009 | −0.0486 | −0.3738 | 0.2752 |
| Location_L2 | 0.3221 | −0.0013 | 0.6857 | 0.1692 | −0.0893 | 0.5136 |
| Location_L3 | 0.0273 | −0.2712 | 0.2865 | −0.0082 | −0.3004 | 0.2436 |
| Location_L4 | 0.0588 | −0.5435 | 0.6616 | 0.0631 | −0.2927 | 0.5398 |
| Behavior_B2 | −0.1824 | −0.5168 | 0.0690 | −0.0940 | −0.3935 | 0.1990 |
| Behavior_B3 | −0.0122 | −0.3403 | 0.2794 | −00258 | −0.3368 | 0.2837 |
| Variance of frailties | 0.5344 | 0.2225 | 0.9447 | 0.5278 | 0.1465 | 1.2583 |
| Covariates | MICE | | | | | |
| | Posterior mean | 2.5% quantile | 97.5% quantile | | | |
| Age_linear | −0.1043 | −0.5104 | 0.2433 | | | |
| Age_quad | −0.0567 | −0.3539 | 0.2395 | | | |
| Location_L2 | 0.1618 | −0.1344 | 0.5570 | | | |
| Location_L3 | 0.0093 | −0.2665 | 0.2621 | | | |
| Location_L4 | 0.0648 | −0.3029 | 0.4223 | | | |
| Behavior_B2 | −0.0512 | −0.3919 | 0.2307 | | | |
| Behavior_B3 | −0.0314 | −0.3252 | 0.2020 | | | |
| Variance of frailties | 0.5177 | 0.1727 | 1.0509 | | | |

showed similar results.

No significant covariate affected the time to reoperation; however, in the SI method, Behavior_B2 was 'nearly significant' since its 95% confidence intervals of regression coefficient was (−0.5168, 0.0690) and the upper bound was near a negative value. The effect was not statistically significant at $p < 0.05$; however, we could assume that patients who had stricturing behavior (B2) tended to have longer time to have a re-operation from CD recurrence compared to patients who had non-stricturing or non-penetrating behavior.

## 6. Discussion

This paper compared five different missing data methods (CC, SI, and three different multiple imputation methods: MI-MICE, MI-MICE-PMM, and MI-AregImpute) in the presence of missing covariates in clustered data. We estimated the regression coefficient through modeling the unknown baseline hazard and frailty density using B-spline in a Bayesian paradigm with incorporating prior distribution. Modeling is flexible and prior information is available when using a Bayesian approach. In the simulation study, performances of the five missing data methods were different depending on the missing rate and the type of covariate. At a missing rate of 5%, the CC analysis showed good performance to estimate the coefficient for a variable that did not have a missing value. However, CC had the worst performance for a covariate with missing values, even when the missing rate was very

small. Single imputation methods showed a better performance than MI methods at a missing rate of 5% for covariates with missing values. This result shows that MI method does not always give better results than SI methods. For categorical variables, aregImpute method produced less biased regression coefficient estimates when 15% or more cases had missing data. However, MICE and MICE-PMM showed good results for continuous variable at a missing rate of 15% (or higher). The performance in handling missing values vary depending on the underlying distribution of covariates and the missing data mechanism. Our simulation results were based on a restricted data setting. Therefore, caution is needed when generalizing the study results. Further study is needed to obtain consistent results under various settings when missing covariates have stratification and time-dependence or skewness. Further simulation is also needed when different distributions of frailty and baseline hazard are assumed

## Acknowledgements

This research was supported by the research grant of the Busan University of Foreign Studies in 2017 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.2017R1C1B5076671). This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2017R1D1A1B03028279).

## References

Brand JPL (1999). *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*, Erasmus University, Rotterdam.

Chen HY and Little RJA (1999). Proportional hazards regression with missing covariates, *Journal of the American Statistical Associations*, **94**, 896–908.

Clayton DG (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65**, 141–152.

Frank H (2010). Hmisc: Miscellaneous library for R statistical software. R package 3.9-0.

Heitjan DF and Little RJA (1991). Multiple imputation for the fatal accident reporting system, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **40**, 13–29.

Lee KY, Yu CS, Lee KY, Cho YB, Park KJ, Choi GS, Yoon SN, and Yoo H (2012). Risk factors for repeat abdominal surgery in Korean patients with Crohn's disease: a multiple-center study of a Korean inflammatory bowel disease study group, *Journal of the Korean Society of Coloproctology*, **28**, 188–194.

Lipsitz SR and Ibrahim JG (1996). Using the EM algorithm for survival data with incomplete categorical covariates, *Lifetime Data Analysis*, **2**, 5–14.

Lipsitz SR and Ibrahim JG (2000). Estimation with correlated censored survival data with missing covariates, *Biostatistics*, **1**, 315–327.

Marshall A, Altman DG, Royston P, and Holder RL (2010). Comparison of techniques for handling missing covariate data within prognostic modeling studies: a simulation study, *BMC Medical Research Methodology*, **10**.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Schenker N and Taylor JMG (1996). Partially parametric techniques for multiple imputation, *Computational Statistics & Data Analysis*, **22**, 425–446.

Sharef E, Strawderman RL, Ruppert D, Cowen M, and Halasyamani L (2010). Bayesian adaptive B-spline estimation in proportional hazard frailty models, *Electronic Journal of Statistics*, **4**, 606–642.

Spiegelhalter DJ, Best NG, Carlin BP, and Van der Linde A (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

van Buuren S (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research*, **16**, 219–242.

van Buuren S, Boshuizen HC, and Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis, *Statistics in Medicine*, **18**, 681–694.

Vaupel JW, Manton KG, and Stallard E (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439–454.

Zhou H and Pepe MS (1995). Auxiliary covariate data in failure time regression, *Biometrika*, **82**, 139–149.

*Received September 13, 2017; Revised December 21, 2017; Accepted December 22, 2017*