Communications for Statistical Applications and Methods
2018, Vol. 25, No. 4, 385–396

https://doi.org/10.29220/CSAM.2018.25.4.385
Print ISSN 2287-7843 / Online ISSN 2383-4757

Multiple imputation for competing risks survival data via pseudo-observations

Seungbong Han^{1,a}, Adin-Cristian Andrei^{b}, Kam-Wah Tsui^{c}

^{a}Department of Applied Statistics, Gachon University, Korea;
^{b}Department of Preventive Medicine, Northwestern University, USA;
^{c}Department of Statistics, University of Wisconsin-Madison, USA

Abstract

Competing risks are commonly encountered in biomedical research. Regression models for competing risks data can be developed based on data routinely collected in hospitals or general practices. However, these data sets usually contain the covariate missing values. To overcome this problem, multiple imputation is often used to fit regression models under a MAR assumption. Here, we introduce a multivariate imputation in a chained equations algorithm to deal with competing risks survival data. Using pseudo-observations, we make use of the available outcome information by accommodating the competing risk structure. Lastly, we illustrate the practical advantages of our approach using simulations and two data examples from a coronary artery disease data and hepatocellular carcinoma data.

Keywords: competing risks, missing data, multiple imputation, pseudo-observations, random forest

1. Introduction

In a competing risks setting, an individual can undergo failure from any of several event types but we observe the first occurring event. A schematic for a single competing risks setting is represented in Figure 1.

In the beginning, an individual is in the initial state 0, but may then experience an event of interest (state 1) or a competing event (state 2). For regression models, several researchers proposed a cumulative incidence function modeling approach assuming independence between subjects (Fine and Gray, 1999; Klein and Andersen, 2005). Logan *et al.* (2011) use marginal models for clustered event times with competing risks based on pseudo-observations (POs). Moreno-Betancur and Latouche (2013) propose a regression model of the cumulative incidence function (CIF) with missing causes of failures using POs. Nicolaie *et al.* (2013) propose dynamic POs method constructed from the prediction probabilities at different landmark times. Kim and Kim (2016) also use POs for the regression modeling of interval censored data and Do and Kim (2017) extend their approach to interval censored data with missing causes of failures. Moreno-Betancur and Latouche (2013) propose a method of the missing cause of failure problem for right-censored data using POs. Ahn and Mendolia (2014) also used POs to compare median survivals for dependent data. POs-based regression is easy to implement using existing software after the POs have been obtained. Therefore it has potential for

[1] Corresponding author: Department of Applied Statistics, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do, 13120, Korea. E-mail: hanseungbong@gachon.ac.kr

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

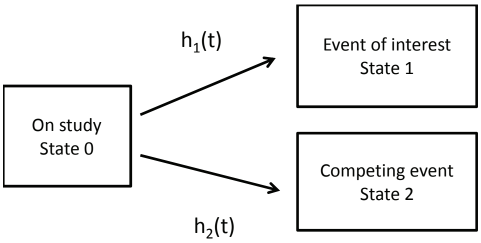

Figure 1: *A schematic for simple competing risks survival data.*

flexible regression modeling. Logan *et al.* (2011) point out several properties of the POs using influence functions by Graw *et al.* (2009). They show thr expected value of POs when the covariate value converges to the cumulative incidence function using an inverse probability of censoring weighting formulation for the cumulative incidence estimate.

We collect registry data sets in hospitals/general practices but these data sets usually contain missing values of the covariates. It is not unusual for more than half the data to be missing for important predictors (Ambler *et al.*, 2005). A naive approach to deal with missing/incomplete covariates is to conduct a complete case analysis (Ambler *et al.*, 2007). In this case, patients with missing values are excluded from the analysis. However, this approach typically results in substantial bias for the regression coefficient estimates and less reliable predictions. Multiple imputation by chained equations (MICE) is another widely used practical approach to deal with missing values (van Buuren *et al.*, 1999). MICE, also known as a full conditional specification, has been used successfully to generate imputed values in survival data analyses (Shah *et al.*, 2014). Shah *et al.* (2014) propose a MICE algorithm using a random forest (RF) method (Breiman, 2001) as the conditional model for imputation. The RF method automatically includes any interaction effect, as well as highly correlated covariates. In addition, the RF method uses bootstrap aggregation of multiple regression trees to reduce the risk of overfitting. Therefore, using an RF has an advantage over MICE methods based on classification and regression trees (CART), as proposed by Burgette and Reiter (2010). For competing risks survival data, we extend the RF-based MICE method using POs. Mogensen and Gerds (2013) also propose a random forest method for competing risks using POs. They first construct the POs and use them as an outcome variable in the random forest model. They compare the predictive performance of their proposed pseudo random forests to that of Cox regression model. Because the model interpretation is essential as well as the prediction error, we use POs as outcome variables in the Fine-Gray model and examine regression coefficient estimates under the scenario of covariate missing, which was not studied before. We compare a standard implementation using a MICE approach based on CART or an RF, and then propose a new version of MICE that uses POs for the competing risks survival data.

The remainder of this paper is organized as follows. We present a PO-based RF MICE (PORF) strategy in Section 2 in detail. Then, in Section 3, we use extensive simulation results to illustrate our method's performance for moderate sample sizes. In Sections 4 and 5, we apply our method to coronary artery disease data and HCC data. Lastly, Section 6 concludes the paper with a discussion of the results.

2. Missing data imputation method for competing risks survival data

Let T_i and C_i be the event time and the censoring time for subject i. Suppose the data consists of n observations and the observation for subject i is $(X_i, \delta_i, \epsilon_i, Z_i)$, where X_i is the observation time as $(T_i \wedge C_i)$ and δ_i is the event indicator as $I(T_i \leq C_i)$. We denote ϵ_i the event type which can be observed if a certain event has occurred before C_i. We also denote Z_i the p-dimensional covariate vector. We assume the censoring time C_i is independent of the event time T_i. For the n observations, $\boldsymbol{Z}$ is the $n \times p$ covariate matrix. We arrange the columns of $\boldsymbol{Z}$ so that $\boldsymbol{Z} = (\boldsymbol{Z}_C, \boldsymbol{Z}_P)$, where $\boldsymbol{Z}_C$ is a completely observed covariate matrix, and $\boldsymbol{Z}_P$ is a partially observed covariate matrix. To conduct the MICE approach, we introduce PO matrix $\boldsymbol{Y}$, where $\boldsymbol{Y}$ is composed of q columns of POs. We obtain the POs at q different time points by using the jackknife method to substitute X_i, δ_i and ϵ_i (Klein and Andersen, 2005). Next, we describe how POs are obtained for the competing risks data. Assume that there are K distinct event types, indexed by $\epsilon_i \in \{1, 2, \ldots, K\}$. It is possible to experience experience failure from any of the K event types; however, we only observe the event time for the first occurring event (or the last follow up time if no failure has occurred). We assume no measurement error and non-informative censoring. For event k, the cause-specific hazard function $h_k(t)$ and overall hazard function $h_.(t)$ are defined as:

$$h_k(t) = \lim_{\Delta t \to 0} \left\{ \frac{\Pr(t < T \leq t + \Delta t, \epsilon = k | T > t)}{\Delta t} \right\}$$

and

$$h_.(t) = \sum_{k=1}^{K} h_k(t).$$

The CIF for event k at time t is obtained by

$$F_k(t) = \Pr(T \leq t, \epsilon = k) = \int_0^t h_k(u) S(u) du,$$

where $S(t) = \exp\{-\int_0^t h_.(u)du\}$ is the overall survival function. Finally, the PO for the i^{th} subject at time t is defined as

$$v_i(t) = n\hat{F}_k(t) - (n-1)\hat{F}_k^{(i)}(t),$$

where $\hat{F}_k(t)$ and $\hat{F}_k^{(i)}(t)$ are the estimated CIFs based on the full sample and a sample of size $n-1$, respectively. In the latter case, we delete the i^{th} observation. To select $q\,(\geq 2)$ distinct time points $0 < t_1 < \cdots < t_q < \infty$ for the PO vector $v_i(t) = (v_i(t_1), \ldots, v_i(t_q))$, we may use percentile values based on the estimated CIF. Andersen and Perme (2010) discuss the choice of the number of time points, q. We select nine time points based on the estimated CIF. These nine points correspond to the t_q values as:

$$t_q = \inf\{t : F_k(t) \geq p\},$$

where $p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$, or 0.9. Originally, the POs were used for regression modeling by Klein and Andersen (2005). We extend their approach for the imputation of missing data using MICE for competing risks survival data. To implement MICE, we rearrange covariate matrix $\boldsymbol{Z}$ such that the complete observation proportion is decreasing from left to right. The MICE procedure is as follows. Let $\psi_1, \psi_2, \ldots, \psi_{p+q}$ constitute the rearranged matrix Ψ. The first variable with the smallest missing variable, say ψ_1, is regressed on the other variables, $\psi_2, \ldots, \psi_{p+q}$. Missing values in ψ_1 are replaced by random draws from a predictive distribution of ψ_1. The next variable with missing values, say ψ_2 is regressed with all other variables, $\psi_1, \psi_3, \ldots, \psi_{p+q}$. Then, missing values in ψ_2 are replaced by random draws from a predictive distribution of ψ_2. This process is repeated for all other variables with missing values, and this cycle constitutes one imputed data set. It is common to use a generalized linear model for the predictive regression model. However, we use the RF algorithm for the regression estimation. The RF algorithm uses bootstrapping sampling such that records with missing values in the dependent variable are imputed by random draws from a normal distribution centered on a conditional mean of the predicted value. We generate several imputed data sets to represent the uncertainty about the missing value. Then we combine the estimates from the multiple imputed data sets using Rubin's rule (Rubin, 1987). When there are B imputed data sets, the combined estimate $\hat{\theta}$ is given by

$$\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b,$$

where $\hat{\theta}_b$ is an estimate of a quantity of interest, such as a regression coefficient. Suppose V_b is the estimated variance of $\hat{\theta}_b$, and V is the average of V_b across the B imputed data sets. Then, the following combined variance accommodates both the within-imputation and the between-imputation variability:

$$\mathrm{var}\left(\hat{\theta}\right) = V + \left(1 + \frac{1}{B}\right)W,$$

where $W = 1/(B-1)\sum_{b=1}^{B}(\hat{\theta}_b - \hat{\theta})^2$. There are several regression models for competing risks data (Aalen *et al.*, 2008). The Fine-Gray method directly models the CIFs (Fine and Gray, 1999). For completeness, we describe the model here. Assume β_k and Z_i are a $1 \times p$ vector of regression coefficients for event k and a $p \times 1$ vector of covariates for subject i, respectively. Then, the Fine-Gray model is given by

$$F_k(t; Z_i) = 1 - \exp\{-\Lambda_0(t) \cdot \exp(\beta_k \cdot Z_i)\},$$

where $i = 1, \ldots, n$, and $\Lambda_0(t)$ is an unspecified non-decreasing baseline. Note that a positive β_k indicates that the CIF increases with Z_i. Furthermore, suppose that C is the censoring time and $G(t)$ is the survival function for C, $\Pr(C \geq t)$. The following score function is proposed to estimate the regression coefficients for the event of interest k:

$$U(\beta_k) = \sum_{i=1}^{n}\int_0^{\infty}\left\{Z_i - \frac{\sum_{j=1}^{n} w_j(s)\tilde{Y}_j(s)Z_j\exp(\beta_k Z_j)}{\sum_{j=1}^{n} w_j(s)\tilde{Y}_j(s)\exp(\beta_k Z_j)}\right\} w_i(s)d\tilde{N}_i(s),$$

where $\tilde{N}_i(t) = I(T_i \leq t, X_{T_i} = k)$, $\tilde{Y}_i(t) = 1 - \tilde{N}_i(t-)$, $r_i = I(C_i \geq \min(T_i, t))$, and $w_i(t) = r_i(t)\{\hat{G}(t)/\hat{G}(\min(t, T_i, C_i))\}$. We can see that $w_j(t)\tilde{Y}_j(t)$ approximates the sub-distribution risk set. Without the competing risks, the weighted score function reduces to the standard score function for the Cox model. In the following simulation section, the PO-based multiple imputation method is compared to the methods of Burgette and Reiter (2010) (CART-based MICE; CRT) and Shah *et al.* (2014) (RF-based MICE; SHAH). CRT is a nonparametric approach to implement multiple imputation via chained equations using sequential regression trees as the conditional models. This method uses the CART imputation engine that can fit interactions, nonlinear relations without data transformations. However, the SHAH method uses the random forest imputation for the multivariate imputation by chained equations.

3. Simulation studies

To evaluate the performance of the proposed method for the competing risk survival data, we compare it against the CRT method and the SHAH method. Here, the **R** packages **tree** and **mice** are used for the simulation comparison (Ripley, 2014; van Buuren and Groothuis-Oudshoorn, 2011). Variables were artificially made under the missing-at-random (MAR) assumption. Three different simulation scenarios are considered, according to the competing event and censoring proportions. Competing risks survival data under the proportional hazards model are generated as follows. For the event of interest in state 1 and the competing event in state 2, we assume cause-specific hazards for subject i, given by:

$$h_1(t|Z_i) = t \cdot \exp\{\beta \cdot Z_i\}$$

and

$$h_2(t|W_i) = t \cdot \exp\{\gamma \cdot W_i\},$$

where the regression parameter $\beta = (1, 1, 1, 1, 0.1, 0.1, 0.1)$, and $\gamma = (1.2, 0.2, 1.2)$ for scenario 1, $\gamma = (1.2, 0.3, 0.8)$ for scenario 2, or $\gamma = (1, 1, 1)$ for scenario 3. The covariate matrix for state 1 is composed of seven dimensional vectors, $Z_i^{'} = (z_1, z_2, z_3, z_4, z_4^2, z_1z_2, z_1z_4)$, where z_1 is drawn from a normal distribution $N(-4, 0.5)$, z_2 is drawn from a binomial distribution Bin(0.5), z_3 is drawn from a uniform distribution $U(-4, -2)$, and z_4 is drawn from $N(-3, 0.5)$. However, the covariate matrix for state 2 is composed of three dimensional vectors, $W_i^{'} = (w_1, w_2, w_3)$, where w_1 is drawn from a normal distribution $N(-6, 1)$, w_2 is drawn from a Weibull distribution Wei(1, 0.5), and w_3 is drawn from a uniform distribution $U(-6, -4)$. First, we simulate the failure time with the all-cause hazard $h.(t|Z_i, W_i) = h_1(t|Z_i) + h_2(t|W_i)$. We then run a binomial experiment for a simulated failure time that decides on the event of interest with probability $h_1(t|Z_i)/h.(t|Z_i, W_i)$ (Beyersmann *et al.*, 2012). In this way, we generate competing risks data and random right-censoring times C. We simulate 1,000 subjects from each design, and delete observations from z_1 through w_3 via a MAR mechanism that depends on z_3, which is completely observed. This leads to around 10% missing values for every variable, except z_3. On average, 60% of the covariate matrix is complete. The γ values change the competing risks proportions so that around 3%, 6%, and 10% of competing events are generated in scenarios 1, 2, and 3. To make the time points for $v_i(t)$ fixed across simulation-runs, we generate 1,000,000 subjects, and select 9 points based on percentiles of the cumulative incidence estimate. We perform multiple imputation using the **randomForest** package, with default settings (Liaw and Wiener, 2002). After the all the covariates are imputed, we fit the competing risks model based on the *crr* function from the **cmprsk** package in **R** (Gray, 2014). Using $B = 20$ or 40 may produce more accurate results in some situations; however, in our case, $B = 10$ produces good results (Graham *et al.*, 2007).

We compared the proposed method (PORF) with the CRT algorithm (Burgette and Reiter, 2010) and the SHAH algorithm (Shah *et al.*, 2014). In addition, we included the complete case analysis results (CC) for the comparison. CRT first sorts the columns of $(Z_i^{'}, W_i^{'})$ to have increasing numbers of missing values. Then it impute missing values in $(Z_i^{'}, W_i^{'})$ with predictive values from the CART model fit using the R function *tree*. In order to yield 10 imputed sets, CRT repeats this step 10 times. However, SHAH uses the predictive values to replace missing values from the random forest fit using the R function *mice*. The *mice* function has wide capabilities so that there are broad imputation engines such as a Bayesian linear regression, a polytomous logistic regression, a linear discriminant analysis and the random forest missing imputation. To implement the standard random forest, we set ‘rf’ for

Table 1: Scenario 1: Averages of the estimated regression coefficients and the standard errors across 1,000 simulation runs; true $\beta = (1, 1, 1, 1, 0.1, 0.1, 0.1)$, competing risk coefficient $\gamma = (1.2, 0.2, 1.2)$

n	Method	$\hat{\beta}_1$	SE_1	$\hat{\beta}_2$	SE_2	$\hat{\beta}_3$	SE_3	$\hat{\beta}_4$	SE_4	$\hat{\beta}_5$	SE_5	$\hat{\beta}_6$	SE_6	$\hat{\beta}_7$	SE_7
300	PORF	0.929	0.535	0.909	1.021	0.934	0.120	0.910	1.131	0.094	0.155	0.084	0.254	0.081	0.165
	CRT	0.499	0.492	1.076	1.016	0.978	0.119	0.074	1.172	0.056	0.162	0.127	0.253	−0.061	0.158
	SHAH	0.575	0.507	0.694	0.997	0.970	0.123	−0.080	1.113	0.029	0.154	0.032	0.247	−0.059	0.159
	CC	1.118	0.979	1.132	1.273	0.987	0.171	1.150	1.833	0.120	0.224	0.108	0.317	0.108	0.317
400	PORF	0.921	0.482	1.002	0.882	0.946	0.105	0.920	0.991	0.098	0.134	0.104	0.219	0.087	0.150
	CRT	0.471	0.439	1.146	0.873	0.958	0.105	0.212	0.999	0.076	0.135	0.141	0.217	−0.060	0.142
	SHAH	0.535	0.448	0.737	0.876	0.950	0.107	0.046	0.988	0.050	0.136	0.039	0.217	−0.062	0.141
	CC	1.106	0.835	1.096	1.085	0.985	0.149	1.161	1.555	0.119	0.189	0.127	0.270	0.111	0.270
500	PORF	0.945	0.439	0.950	0.783	0.954	0.096	0.960	0.886	0.117	0.120	0.092	0.195	0.091	0.136
	CRT	0.469	0.397	1.070	0.777	0.946	0.096	0.257	0.892	0.085	0.121	0.125	0.193	−0.064	0.128
	SHAH	0.552	0.416	0.654	0.802	0.939	0.098	0.142	0.909	0.064	0.126	0.022	0.199	−0.060	0.130
	CC	0.981	0.748	1.070	0.971	0.979	0.133	1.107	1.391	0.117	0.167	0.122	0.242	0.101	0.242

PORF = PO-based RF MICE; CRT = CART-based MICE; SHAH = RF-based MICE; CC = complete case analysis results. PO = pseudo-observation; RF = random forest; MICE = multiple imputation by chained equations; CART = classification and regression trees.

Table 2: Scenario 2: Averages of the estimated regression coefficients and the standard errors across 1,000 simulation runs; true $\beta = (1, 1, 1, 1, 0.1, 0.1, 0.1)$, competing risk coefficient $\gamma = (1.2, 0.3, 0.8)$

n	Method	$\hat{\beta}_1$	SE_1	$\hat{\beta}_2$	SE_2	$\hat{\beta}_3$	SE_3	$\hat{\beta}_4$	SE_4	$\hat{\beta}_5$	SE_5	$\hat{\beta}_6$	SE_6	$\hat{\beta}_7$	SE_7
300	PORF	0.904	0.560	0.929	1.074	0.901	0.126	0.918	1.169	0.098	0.160	0.101	0.266	0.062	0.174
	CRT	0.420	0.518	1.050	1.068	0.873	0.126	0.237	1.219	0.078	0.168	0.132	0.264	−0.058	0.167
	SHAH	0.452	0.506	0.678	1.061	0.867	0.129	−0.007	1.167	0.044	0.164	0.039	0.262	−0.068	0.161
	CC	0.901	1.018	1.086	1.328	0.893	0.177	1.055	1.913	0.110	0.234	0.116	0.329	0.096	0.330
400	PORF	0.914	0.504	0.921	0.938	0.920	0.109	0.915	1.044	0.093	0.142	0.099	0.232	0.075	0.157
	CRT	0.440	0.456	1.057	0.931	0.863	0.109	0.212	1.056	0.067	0.144	0.133	0.230	−0.047	0.148
	SHAH	0.473	0.453	0.677	0.937	0.856	0.112	0.025	1.036	0.043	0.146	0.039	0.231	−0.057	0.143
	CC	0.884	0.876	1.071	1.144	0.887	0.154	0.995	1.638	0.102	0.198	0.135	0.283	0.096	0.283
500	PORF	0.922	0.462	0.921	0.839	0.932	0.099	0.922	0.948	0.103	0.129	0.100	0.207	0.081	0.143
	CRT	0.413	0.410	1.075	0.831	0.853	0.098	0.212	0.949	0.073	0.129	0.139	0.205	−0.056	0.133
	SHAH	0.496	0.414	0.684	0.845	0.846	0.100	0.099	0.950	0.051	0.134	0.042	0.209	−0.051	0.131
	CC	0.889	0.784	1.038	1.028	0.874	0.138	1.001	1.461	0.102	0.175	0.129	0.254	0.096	0.253

PORF = PO-based RF MICE; CRT = CART-based MICE; SHAH = RF-based MICE; CC = complete case analysis results. PO = pseudo-observation; RF = random forest; MICE = multiple imputation by chained equations; CART = classification and regression trees.

the method options. Tables 1–3 show the simulation results for the three scenarios, respectively. The simulation results are summarized across 1,000 Monte Carlo replicates.

Table 1 displays the averages of the estimated regression coefficient and the standard errors. For the main effect terms, the CRT and the SHAH methods significantly underestimate the true value of β. In contrast, the proposed PORF method produces regression coefficients with less bias than the CRT and SHAH methods. Note that the estimates of β_3 are relatively similar compared to other coefficients in Table 1. We think this is due to its significance. When the competing risk proportion increases, the PORF performs better than others (Table 2 and Table 3). Other factors such as the competing risk proportion and the covariate significance might affect the coefficient estimation. The bias in the CRT and SHAH algorithms does not decrease when the sample size increases to 400 and 500. In addition, for the quadratic and interaction terms, the results using the proposed method are less biased than those of the CRT and SHAH methods. The reason for the heavy bias in CART and SHAH might come from the fact that these methods are originally designed for standard survival. We also note that

Table 3: Scenario 3: Averages of the estimated regression coefficients and the standard errors across 1,000 simulation runs; true β = (1, 1, 1, 1, 0.1, 0.1, 0.1), competing risk coefficient γ = (1, 1, 1)

n	Method	$\hat{\beta}_1$	SE$_1$	$\hat{\beta}_2$	SE$_2$	$\hat{\beta}_3$	SE$_3$	$\hat{\beta}_4$	SE$_4$	$\hat{\beta}_5$	SE$_5$	$\hat{\beta}_6$	SE$_6$	$\hat{\beta}_7$	SE$_7$
300	PORF	0.915	0.582	0.946	1.105	0.922	0.127	0.914	1.216	0.111	0.166	0.110	0.273	0.083	0.181
	CRT	0.365	0.543	1.073	1.098	0.832	0.127	0.230	1.269	0.083	0.174	0.143	0.272	−0.066	0.176
	SHAH	0.415	0.523	0.672	1.091	0.827	0.130	0.024	1.217	0.053	0.172	0.042	0.270	−0.071	0.167
	CC	0.895	1.041	0.988	1.358	0.869	0.179	1.031	1.954	0.106	0.239	0.116	0.337	0.098	0.337
400	PORF	0.914	0.524	0.947	0.962	0.914	0.111	0.918	1.078	0.090	0.147	0.111	0.238	0.084	0.162
	CRT	0.366	0.473	1.084	0.955	0.826	0.111	0.206	1.099	0.075	0.150	0.146	0.236	−0.061	0.153
	SHAH	0.429	0.465	0.676	0.944	0.820	0.113	−0.009	1.080	0.041	0.153	0.044	0.233	−0.062	0.147
	CC	0.844	0.897	1.046	1.172	0.852	0.156	0.915	1.682	0.093	0.203	0.133	0.290	0.090	0.290
500	PORF	0.927	0.475	0.903	0.858	0.912	0.099	0.923	0.976	0.105	0.132	0.100	0.212	0.087	0.147
	CRT	0.386	0.422	1.038	0.852	0.821	0.099	0.205	0.973	0.072	0.132	0.134	0.210	−0.057	0.136
	SHAH	0.444	0.421	0.667	0.861	0.816	0.101	0.068	0.987	0.050	0.139	0.041	0.213	−0.058	0.133
	CC	0.850	0.803	0.985	1.052	0.843	0.139	0.943	1.499	0.096	0.180	0.12	0.26	0.089	0.259

PORF = PO-based RF MICE; CRT = CART-based MICE; SHAH = RF-based MICE; CC = complete case analysis results. PO = pseudo-observation; RF = random forest; MICE = multiple imputation by chained equations; CART = classification and regression trees.

the bias of CC is comparable to the PORF method. However, the average of standard errors has almost doubled compared to the PORF. Dropping the incomplete cases in the analysis translates directly to a loss in sample size to estimate the regression coefficients. Therefore, the complete case analysis produces a loss of efficiency.

4. The coronary artery disease data example

Coronary artery disease (CAD) is the leading cause of death in the US and Europe. It is one of the most common types of heart disease and occurs when the arteries become narrowed. The build-up of cholesterol and other materials such as plaque on inner walls is the main cause. Asan Medical Center has constructed a revascularization registry to study left main or multivessel CAD (Seung *et al.*, 2008). They have compared the treatment effect of a percutaneous coronary intervention (PCI) with coronary-artery bypass grafting (CABG) regarding long-term outcomes. Long-term adverse outcomes were measured in terms of death, myocardial infarction (MI), and stroke or target-vessel revascularization. A combined endpoint of death, MI, stroke, and the revascularization was considered a primary endpoint in the previous study; however, it is also possible to investigate the effect of the coronary stents along with other clinical, angiographic or procedural variables for cardiac-specific death. We included 5,775 patients in the analysis and several clinical (age, comorbidity, hemodynamic status, clinical presentations, and prior history of PCI or CABG) and angiographic factors (coronary anatomy, disease extents, and procedural complexities) were considered as possible factors for cardiac-specific death. The covariate variables are age (years), gender, body mass index (BMI), hypertension (yes/no), diabetes (yes/no), current smoker (yes/no), hyperlipidemia (yes/no), PCI intervention (yes/no), prior MI (PMI; yes/no), prior PCI (PPCI; yes/no), prior congestive heart failure (PCHF; yes/no), chronic lung disease (CLD; yes/no), previous stroke (PST; yes/no), peripheral vascular disease (PVD; yes/no), renal dysfunction (RD; yes/no), electrocardiographic finding (ELF; sinus rhythm vs atrial fibrillation + others), acute coronary syndrome (ACS; yes/no), ejection fraction (EF; %), left main (LM) disease (yes/no), proximal left anterior descending artery (PLAD) disease (yes/no), right CAD (yes/no), bifurcation lesion (BL; yes/no), restenotic lesion (RL; yes/no), total occlusion (TO; ≥ 1 vs < 1), EuroSCORE, and SYNTAX Score. After the surgery, 446 patients and 294 patients died of cardiac related and non-cardiac related causes, respectively during the follow-up.

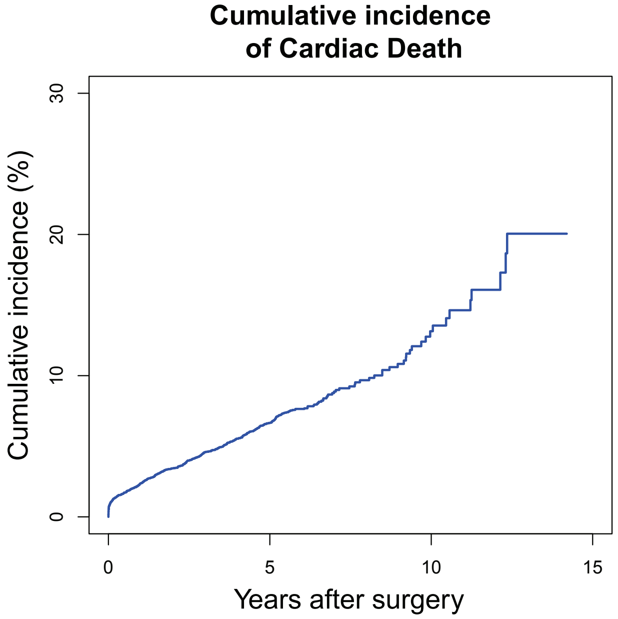

Figure 2: *Cumulative incidence function for cardiac death.*

Figure 2 shows the cumulative incidence rate of the cardiac death. The cumulative incidence funtion is estimated using the **R** function *cuminc* from the **cmprsk** package. It seems the risk increases gradually to 10 years and starts to ascend more steeply. Among the covariate the following variables have missing values : BMI(1.2%), RF(14.6%), ACS(12.2%), EuroSCORE (13.1%), SYNTAX score(39.6%), PLAD(13.9%), BL(10.4%), and TO(8.6%). Only 3442 patients have complete data on all covariates. We apply PORF, CRT, and SHAH to the data to deal with the missing values. Then we fit both univariate and multivariate Fine-Gray regression models. To construct the POs, we use nine time points based on the CIF of the cardiac death. Supplementary Table 1 shows the results of the univariate analysis. Several factors such as age, diabetes, prior MI, prior congestive heart failure, previous stroke, peripheral vascular disease, renal dysfunction, atrial fibrillation, and bifurcation lesion seem to increase the cardiac mortality significantly. However, factors such as large BMI, the PCI and the large ejection fraction level tend to decrease the cardiac mortality risk. In the multivariate regression (Table 4), the final model based on the PORF method is similar with the final model using CRT because the missing proportion is low. Note that the SYNTAX score is not statistically significant in the analysis result based on the SHAH method.

5. The hepatocellular carcinoma data example

We apply the PORF method to a retrospective study of hepatocellular carcinoma data as the second example. The data comprise 20 covariate variables measured on 525 patients who experienced curative hepatectomies at the Asan Medical Center in Korea from 2000 to 2006. The primary study endpoint is HCC-specific death after hepatic resection surgery, but patients can also die of non-cancer-related causes. Thus, our study uses regression modeling to predict post-hepatectomy outcomes. All patients underwent magnetic resonance imaging (MRI), computed tomography (CT), chest CT, and bone scintigraphy, along with serum hepatitis markers. Therefore, we could collect several covari-

Table 4: Multivariate analysis results for the coronary artery disease data

Covariate	PORF			CRT			SHAH		
	$\hat{\beta}$	SE	P	$\hat{\beta}$	SE	P	$\hat{\beta}$	SE	P
Age, year	0.031	0.007	<0.001	0.034	0.007	<0.001	0.035	0.007	<0.001
Diabetes									
yes	0.237	0.098	0.016	0.241	0.098	0.014	0.257	0.098	0.009
RD									
yes	0.584	0.187	0.002	0.605	0.188	0.001	0.599	0.189	0.002
ELF									
AF + others	0.552	0.200	0.006	0.566	0.198	0.004	0.516	0.201	0.010
LM disease									
yes	0.646	0.121	0.001	0.644	0.121	0.001	0.762	0.111	<0.001
EF, %	−0.023	0.004	<0.001	−0.025	0.004	<0.001	−0.025	0.004	<0.001
SYNTAX Score	0.011	0.004	0.010	0.011	0.004	0.007			
EuroSCORE	0.128	0.027	<0.001	0.112	0.027	<0.001	0.120	0.027	<0.001

PORF = PO-based RF MICE; CRT = CART-based MICE; SHAH = RF-based MICE; RD = renal dysfunction; ELF = electrocardiographic finding; EF = ejection fraction; LM = left main; PO = pseudo-observation; RF = random forest; MICE = multiple imputation by chained equations; CART = classification and regression trees.

ates with regard to histological and surgical information, as well as clinical information (Shim *et al.*, 2012). Of the 20 covariates, 11 covariates have missing values between 5% and 38% of their values. The acquired covariate variables are: age (years), gender, body mass index (BMI; < 23 vs ≥ 23), Child-Pugh class (A/B), etiology (hepatitis B virus vs hepatitis C virus + others), indocyanine green retention rate at 15 minutes (ICG R15; < 14% vs ≥ 14%), serum aspartate aminotransferase (AST; IU/L), serum alanine aminotransferase (ALT; IU/L), liver cirrhosis (yes/no), tumor size (cm), alpha-fetoprotein (AFP; ng/mL), number of tumors, microvascular invasion (yes/no), capsular invasion (yes/no), microsatellite lesion (yes/no), Edmonson grade (I or II vs III or IV), American Joint Committee on Cancer stage (AJCC; I vs II + IIIA), resection type (major/minor), resection margin width (< 10 mm vs ≥ 10 mm), and red cell transfusion (yes/no). After the surgery, 9 patients and 142 patients died of non-cancer-related causes and the HCC-related causes, respectively, whereas 374 patients were censored before experiencing HCC-specific death. Our analysis explores possible risk factors related to HCC-specific death. Supplementary Figure 1 shows the cumulative incidence rate of HCC-specific death. After about 3 years, the cumulative incidence rate increases beyond 10%, and increases to 40% at around 10 years.

We fit both univariate and multivariate Fine-Gray regression models. With regard to the covariate variables, the continuous variables of AFP, tumor size, AST, and ALT are severely skewed. Thus, we use a log transformation for these variables to obtain more stable parameter estimates. Many variables have complete records, but 11 have missing values between 5% and 38%. The missing rates are mostly modest; however, they are scattered among the variables so that only 132 patients have complete data on all variables. Assuming a MAR mechanism, we create $B = 10$ complete data sets using the PORF method. We order variables according to the number of values they are missing, from smallest to largest, as discussed in Section 2. As in the simulation study, to construct the POs, we use nine time points based on the CIF of the HCC-related death. Supplementary Table 2 shows the results of the univariate analysis. The results based on the CRT and the SHAH methods are also included in the table. First, it seems that the log-transformed tumor size, male gender, and microvascular invasion status increase the HCC-specific mortality significantly. However, a large BMI (≥ 23) and the Child-Pugh class (A) tend to decrease HCC-specific mortality risk. While the PORF method produces similar results to those of the CRT and SHAH overall, the AJCC stage is only significant in the PORF method. The final model based on the PORF method is similar to the final model using CRT in the

Table 5: Multivariate analysis results for the HCC data

Covariate	PORF			CRT			SHAH		
	$\hat{\beta}$	SE	P	$\hat{\beta}$	SE	P	$\hat{\beta}$	SE	P
Log tumor size	0.421	0.153	0.006	0.420	0.150	0.005	0.471	0.151	0.002
BMI									
≥ 23	−0.451	0.176	0.010	−0.451	0.176	0.010	−0.448	0.176	0.011
Child-Pugh class									
A	−1.215	0.396	0.002	−1.146	0.355	0.001			
AJCC stage									
II and IIIA	0.628	0.270	0.020	0.637	0.260	0.014	0.619	0.273	0.023
MIC invasion									
yes	0.688	0.255	0.007	0.706	0.245	0.004	0.610	0.256	0.017

PORF = PO-based RF MICE; CRT = CART-based MICE; SHAH = RF-based MICE; BMI = body mass index; AJCC = American Joint Committee on Cancer; MIC = microvascular; PO = pseudo-observation; RF = random forest; MICE = multiple imputation by chained equations; CART = classification and regression trees.

multivariate regression (Table 5). The log-transformed tumor size, low BMI (< 23), Child-Pugh class (B), AJCC stage (II and IIIA), and microvascular invasion are risk factors for HCC-specific mortality. Note that the AJCC stage is not statistically significant in the univariate regression based on the CRT method; however, it is significant in the final multivariate model. Thus, this variable could be omitted when using only the CRT method. However, the SHAH method does not show the significance of the Child-Pugh class.

6. Discussion

Missing data for covariates are inevitable when patient data are collected from various aspects. If there is more than one competing event and some of event types are rare, we simply combine them into a composite outcome. However, ignoring competing risks may lead to misleading or even erroneous results that could obstruct the understanding of survival trends. MICE is an increasingly popular method for multiple imputation (Royston and White, 2011). Here, we propose another MICE algorithm using POs for competing risks data. This method uses the RF as an imputation engine so that it can flexibly deal with interactions, nonlinear relations, and complex distributions without parametric assumptions. Using the POs method, we make better use of the available outcome information by accommodating the competing risks structure. The proposed method results in less biased estimates than those of the CRT and SHAH methods in the simulation studies. The results based on real data show that there is marginal difference among different imputation methods. Sample sizes are 5775 and 525 for the CAD study and HCC study, respectively. Overall, estimated standard errors are quite small due to the large sample size for the CAD study. Therefore, all three methods produce similar results. For the HCC study, the missing rates are 19%, 7%, and 8% for the Child-Pugh class, AJCC stage, and MIC invasion variables in the final model. These small missing rates might attenuate differences among the methods. Additional research is needed in future work to deal with missing outcome variables and to extend the proposed method into a multi-state model setting.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2017R1C1B 1006717).

References

Aalen O, Borgan Ø, and Gjessing H (2008). *Survival and Event History Analysis*, Springer, New York.

Ahn KW and Mendolia F (2014). Pseudo-value approach for comparing survival medians for dependent data, *Statistics in Medicine*, **33**, 1531–1538.

Andersen PK and Perme MP (2010). Pseudo-observations in survival analysis, *Statistical Methods in Medical Research*, **19**, 71–99.

Ambler G, Omar RZ, Royston P, Kinsman R, Keogh BE, and Taylor KM (2005). Generic, simple risk stratification model for heart valve surgery, *Circulation*, **112**, 224–231.

Ambler G, Omar RZ, and Royston P (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome, *Statistical Methods in Medical Research*, **16**, 277–298.

Moreno-Betancur M and Latouche A (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values, *Statistics in Medicine*, **32**, 3206–3223.

Beyersmann J, Allignol A, and Schumacher M (2012). *Competing Risks and Multistate Models with R*, Springer-Verlag New York, Chapter 3, 45–50.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Burgette LF and Reiter JP (2010). Multiple imputation for missing data via sequential regression trees, *American Journal of Epidemiology*, **172**, 1070–1076.

Do G and Kim YJ (2017). Analysis of interval censored competing risk data with missing causes of failure using pseudo values approach, *Journal of Statistical Computation and Simulation*, **87**, 631–639.

Fine J and Gray R (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.

Graham JW, Olchowski AE, and Gilreath TD (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory, *Prevention Science*, **8**, 206–213.

Gray B (2014). cmprsk: Subdistribution Analysis of Competing Risks, *R package version 2.2-7.* http://CRAN.R-project.org/package=cmprsk

Graw F, Gerds TA, and Schumacher M (2009). On pseudo-values for regression analysis in competing risks models, *Lifetime Data Analysis*, **15**, 241–255.

Kim S and Kim YJ (2016). Regression analysis of interval censored competing risk data using a pseudo-value approach, *Communications for Statistical Applications and Methods*, **23**, 555–562.

Klein JP and Andersen PK (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function, *Biometrics*, **61**, 223–229.

Liaw A and Wiener M (2002). Classification and regression by randomForest, *R News*, **2**, 18–22.

Logan BR, Zhang MJ, and Klein JP (2011). Marginal models for clustered time to event data with competing risks using pseudovalues, *Biometrics*, **67**, 1–7.

Mogensen UB and Gerds TA (2013). A random forest approach for competing risks based on pseudovalues, *Statistics in Medicine*, **32**, 3102–3114.

Nicolaie MA, van Houwelingen JC, de Witte TM, and Putter H (2013). Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks, *Biometrics*, **69**, 1043–1052.

Ripley B (2014). tree: Classification and regression trees. R package version 1.0-35, from: http://CRAN.R-project.org/package=tree

Royston P and White IR (2011). Multiple imputation by chained equations (MICE): implementation in Stata, *Journal of Statistical Software*, **45**, 1–20.

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

Shah AD, Bartlett JW, Carpenter J, Nicholas O, and Hemingway H (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study, *American Journal of Epidemiology*, **179**, 764–774.

Seung KB, Park DW, Kim YH *et al.* (2008). Stents versus coronary-artery bypass grafting for left main coronary artery disease, *The New England Journal of Medicine*, **358**, 1781–1792.

Shim JH, Yoon DL, Han S, *et al.* (2012). Is Serum Alpha-Fetoprotein useful for predicting recurrence and mortality specific to hepatocellular carcinoma after hepatectomy? A test based on propensity scores and competing risks analysis, *Annals of Surgical Oncology*, **19**, 3687–3696.

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

van Buuren S and Groothuis-Oudshoorn K (2011). mice: multivariate imputation by chained equations in R, *Journal of Statistical Software*, **45**, 1–67.

Received March 8, 2018; Revised April 5, 2018; Accepted May 16, 2018