Communications for Statistical Applications and Methods
2023, Vol. 30, No. 1, 1–19

https://doi.org/10.29220/CSAM.2023.30.1.001
Print ISSN 2287-7843 / Online ISSN 2383-4757

A case study of competing risk analysis in the presence of missing data

Limei Zhou[a], Peter C. Austin[1, a,b,c], Husam Abdel-Qadir[a,d]

[a]Institute for Clinical Evaluative Sciences (ICES), Toronto, Ontario, Canada;
[b,]Institute of Health Management, Policy and Evaluation, University of Toronto, Ontario, Canada
[c]Sunnybrook Research Institute, Toronto, Ontario, Canada;
[d]Division of Cardiology and Department of Medicine, Women's College Hospital, Toronto, Ontario, Canada

Abstract

Observational data with missing or incomplete data are common in biomedical research. Multiple imputation is an effective approach to handle missing data with the ability to decrease bias while increasing statistical power and efficiency. In recent years propensity score (PS) matching has been increasingly used in observational studies to estimate treatment effect as it can reduce confounding due to measured baseline covariates. In this paper, we describe in detail approaches to competing risk analysis in the setting of incomplete observational data when using PS matching. First, we used multiple imputation to impute several missing variables simultaneously, then conducted propensity-score matching to match statin-exposed patients with those unexposed. Afterwards, we assessed the effect of statin exposure on the risk of heart failure-related hospitalizations or emergency visits by estimating both relative and absolute effects. Collectively, we provided a general methodological framework to assess treatment effect in incomplete observational data. In addition, we presented a practical approach to produce overall cumulative incidence function (CIF) based on estimates from multiple imputed and PS-matched samples.

Keywords: missing data, multiple imputation, propensity score matching, competing risk analysis, cumulative incidence function

1. Introduction

Observational data are commonly used in health and epidemiologic studies with advantages of lower cost of data collection in more generalizable settings and the ability to detect rare adverse events (Boyko, 2013). However, due to systematic difference in baseline characteristics between exposed and unexposed subjects, statistical methods often need to be used to account for these differences to ensure valid statistical inference. In recent years, propensity score (PS) matching has been increasingly used in observational studies to mitigate the effects of confounding. With this approach, exposed subjects are matched with unexposed counterparts on the estimated PS, which is the predicted probability of exposure conditional on measured baseline covariates. If successful, the matched subjects have similar PS and the corresponding difference in measured baseline characteristics is reduced. This allows the effects of treatment to be estimated by directly comparing outcomes between exposed and unexposed subjects in the matched samples (Austin, 2011).

[1] Corresponding author: Institute of Health Management, Policy and Evaluation, University of Toronto, ICES, 2075 Bayview Avenue, Toronto, Ontario, M4N 3M5, Canada. E-mail: peter.austin@ices.on.ca

Published 31 January 2023 / journal homepage: http://csam.or.kr

Competing risks arise when subjects can experience different types of outcomes and the occurrence of one type of outcome will preclude the occurrence of others. For example, death due to non-cardiovascular causes is a competing risk for cardiovascular death, as a subject who dies of cancer is no longer at risk of cardiovascular death. The cumulative incidence function (CIF) is a valid estimate of the cumulative incidence of a cause-specific event over time in the presence of competing risks. In addition, the effects of a given set of covariates on the event of interest can be modelled using either a cause-specific hazard model or a subdistribution hazard model, depending on individual research questions (Austin *et al.*, 2016). With the application of PS-matching in the observational competing risk data, a complete examination of treatment effect on cause-specific events can be conducted in the matched exposed and unexposed subjects. In doing so, cause-specific hazard regression models are suitable for relative measures of treatment effects, while tests for differences in absolute treatment effects can be obtained by comparing CIFs using a marginal subdistribution hazard model. When applied to propensity-score matched samples, both models should use a robust variance estimator to account for within-pair clustering of outcomes in the matched samples (Austin and Fine, 2019).

Missing data are a pervasive problem in health research (Nguyen *et al.*, 2017). Several approaches have been developed to handle missing data including complete case analysis, single imputation, maximum likelihood estimation, Bayesian estimation and multiple imputation (MI) (Liu and De, 2015). Among them, MI has gained popularity in in recent years. It fills missing values with multiple plausible values, which explicitly incorporates the uncertainty of missing data (Austin *et al.*, 2021). Furthermore, this approach is computationally straightforward, relatively easy to apply, and increasingly available in standard statistical software. Two iterative methods are available for doing multiple imputation including the joint modeling (JM) and the fully conditional specification (FCS) (Liu and De, 2015). Joint modeling assumes joint multivariate normality of all variables which may be inappropriate for categorical variables and skewed continuous variables, whereas FCS offers more flexibility by sequentially fitting suitable regression models for each incomplete variable, conditional on all other variables in the imputation model (Liu and De, 2015; Huque *et al.*, 2018). Once multiple imputed datasets are generated, standard statistical analysis is conducted in each imputed dataset, and the resulting coefficients and standard errors are then pooled and integrated using Rubin's rule to generate final overall estimates (Little and Rubin, 1987).

In this case study, we used observational competing risk data with incomplete variables to estimate the effect of exposure to statins on heart failure (HF)-related hospital presentations in a cohort of breast cancer patients receiving trastuzumab-based chemotherapy. We aimed to provide a general methodological framework to assess treatment effect in incomplete observational data, with a focus on solving a challenge often seen in producing overall CIFs from individual CIFs derived from multiple imputed and PS-matched samples.

2. Methodology

2.1. Study cohort

The study cohort was previously described in detail (Abdel-Qadir *et al.*, 2021). Briefly, 1,371 women who received trastuzumab chemotherapy within a year of being diagnosed with early breast cancer at age $\geq$ 66 years were identified from health administrative databases with index chemotherapy starting dates between January 1, 2007 and December 31, 2017. The objective of the study was to determine the effect of statin exposure on the risk of heart failure (HF)-related hospital presentations (hospitalizations or emergency department visits). The administrative data at the level of Ontario population are housed at ICES (formerly the Institute for Clinical Evaluative Sciences), Ontario, Canada.

Women were defined as statin-exposed if they had at least 2 statin prescriptions within one year prior to index chemotherapy starting date. The time-to-event outcome was defined as hospital presentations due to HF, and the follow-up time was defined as time in days from index date to a hospital presentation. Death was treated as a competing risk. Patients who were event-free were censored at the end of follow-up on December 31, 2018.

Variables included in the PS model were age, rural residence, neighborhood income quintile, year of diagnosis, stage, left-sided disease, the Charlson score and past medical history within 5 years prior to index date (hypertension, diabetes, chronic obstructive pulmonary disease, chronic kidney disease, atrial fibrillation, acute myocardial infarction (AMI), ischemic heart disease without prior AMI, peripheral vascular disease, non-statin lipid lowering therapy, angiotensin antagonists and beta-blockers). It was important to account for low-density lipoprotein (LDL) level since it is associated with both statin exposure and with the risk of cardiovascular disease. However, statins were hypothesized to reduce the risk of HF after chemotherapy independent of their effect on LDL levels.

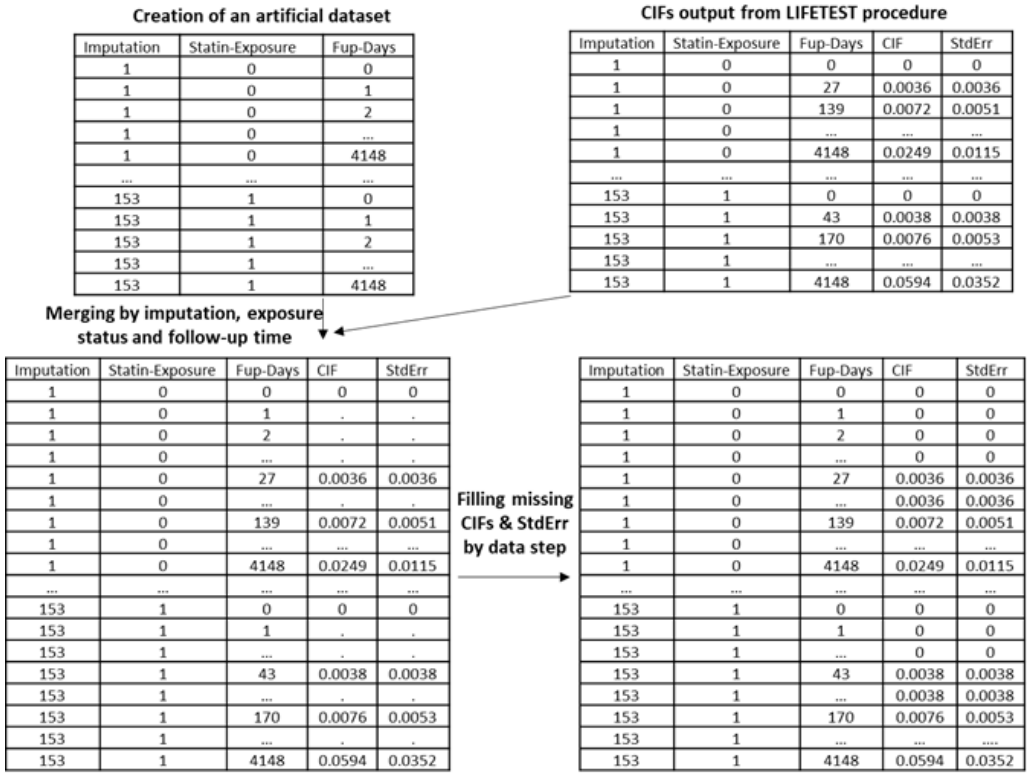
3. Statistical analysis

3.1. Imputation

Four variables were subject to missingness: Rural residence, neighborhood income quintile, left-sided disease and LDL. Among them, the frequencies of missing data in rural residence, neighborhood income quintile and laterality were minimal, 7 in total. However, 693 out of 1,371 (50.5%) patients had missing LDL. Overall, 695 out of 1,371 (50.7%) patients had at least one variable with missing data.

Given the high prevalence of missing data and the reduction in statistical power that would be observed if doing a complete case analysis (in addition to any biases introduced by a complete case analysis), we used the multivariate imputation using chained equations (MICE) algorithm with the approach of fully conditional specification (FCS) implemented in the SAS Proc MI to impute the missing values in the above four explanatory variables. Logistic regression was used to impute categorical variables (rural residence, income quintile and left-sided disease), while predictive mean matching (PMM) was used for the continuous variable LDL. We ordered the FCS model statements based on the percentages of missing values in each variable with the variable having the least missing values at the beginning. Each of the incomplete variables was included in each other's imputation models. Last, in each FCS model statement, the sequence of covariates was ordered as well, with outcome variables (including both HF-related hospital presentations and follow-up time) at the beginning, followed by fully observed explanatory variables and then by variables with missing values in ascending order. The variables in the MI models were described in detail previously (Abdel-Qadir *et al.*, 2021). In this way, 153 datasets were imputed, three times the number of overall missing percentages in the cohort. We initially used White *et al.* (2011)'s suggestion that the number of imputed datasets be equal to the percentage of subjects with missing data (51%). However, we found greater stability in results when using 3 times this number (153 imputed datasets).

We then examined the quality of imputed LDL values both graphically and numerically as the missing percentages for rural residence, income quintile and laterality were minimal. We compared the kernel density estimates of the imputed values to the observed values. In addition, we calculated absolute difference in means between the observed and imputed values as well as a ratio of variances of the observed and imputed values. Those outliers with absolute difference greater than 2 standard deviations, or those with ratio of variance less than 0.5 or greater than 2 were flagged (Stuart *et al.*, 2009).



Imputation	Statin-Exposure	Fup-Days
1	0	0
1	0	1
1	0	2
1	0	...
1	0	4148
...	...	...
153	1	0
153	1	1
153	1	2
153	1	...
153	1	4148

Imputation	Statin-Exposure	Fup-Days	CIF	StdErr
1	0	0	0	0
1	0	27	0.0036	0.0036
1	0	139	0.0072	0.0051
1	0	...	...	...
1	0	4148	0.0249	0.0115
...	...	...	...	...
153	1	0	0	0
153	1	43	0.0038	0.0038
153	1	170	0.0076	0.0053
153	1	...	...	...
153	1	4148	0.0594	0.0352

Imputation	Statin-Exposure	Fup-Days	CIF	StdErr
1	0	0	0	0
1	0	1	.	.
1	0	2	.	.
1	0	...	.	.
1	0	27	0.0036	0.0036
1	0	...	.	.
1	0	139	0.0072	0.0051
1	0	...	...	...
1	0	4148	0.0249	0.0115
...	...	...	...	...
153	1	0	0	0
153	1	1	.	.
153	1	...	.	.
153	1	43	0.0038	0.0038
153	1	...	.	.
153	1	170	0.0076	0.0053
153	1	...	.	.
153	1	4148	0.0594	0.0352

Imputation	Statin-Exposure	Fup-Days	CIF	StdErr
1	0	0	0	0
1	0	1	0	0
1	0	2	0	0
1	0	...	0	0
1	0	27	0.0036	0.0036
1	0	...	0.0036	0.0036
1	0	139	0.0072	0.0051
1	0	...	...	...
1	0	4148	0.0249	0.0115
...	...	...	...	...
153	1	0	0	0
153	1	1	0	0
153	1	...	0	0
153	1	43	0.0038	0.0038
153	1	...	0.0038	0.0038
153	1	170	0.0076	0.0053
153	1	...	...	....
153	1	4148	0.0594	0.0352

Figure 1: *Flow diagram for generation of a dataset with daily CIFs for the entire follow up period. Numbers for observed event time, CIFs and standard errors are for illustration only.*

Subsequently, in each of the imputed datasets, we matched each statin-exposed patient with an unexposed patient at the ratio of 1:1 using greedy nearest neighbor matching within a caliper distance of 0.2 of the standard deviation of the logit of the PS, which was computed using a logistic regression model with statin exposure as the outcome. The detailed modeling including covariates information was described previously (Abdel-Qadir *et al.*, 2021).

Next, we computed CIFs in each exposure group within each of the 153 matched samples and estimated cause-specific hazard ratios accounting for death as a competing risk. The overall estimates were then computed by pooling the results across the 153 imputed samples using the SAS MIanalyze procedure in which Rubin's rules were implemented. Specifically, to quantify relative effects of statin exposure on the HF-related hospital presentations, we used cause-specific hazard model with robust variance estimator to account for within-pair clustering, whereas for quantification of absolute treatment effect of statin exposure, we used the SAS LIFETEST procedure with the option "eventcode" to estimate CIFs for the duration of follow-up time in the matched statin exposed and unexposed patients, respectively. The significance level for the comparison of CIFs between these two groups was conducted using a marginal Fine-Gray subdistribution hazard model (statin exposure as the only covariate) with robust variance estimator as well to account for clustering.

To produce CIF graphs, we created a grid from one to the maximal follow-up time in increment

of one day which is $12 * 365$. Then CIF values were estimated at each value on the grid in each sample. When a CIF was missing for a specific time, the value was then estimated using the previously observed event time as CIF is a step function. Specifically, first we generated an artificial dataset by defining a grid from one to $12 * 365$ in increment of one day, which was then duplicated in each of the statin exposure groups across the 153 imputed samples. In our case, the variable statin exposure was defined as 0 or 1 with two levels, the maximal follow-up time was up to 12 years and 153 complete datasets were constructed. As such, this artificial dataset consisted of 1,340,280 ($2 * 12 * 365 * 153$) observations with three variables including imputation numbers, statin exposure status and follow-up time in days. Subsequently, we used the previously described LIFETEST procedure to estimate and output the CIFs and corresponding standard errors (STDERR) at each observed event time in days which was renamed as follow-up time, we then merged these two datasets by the above three variables. We then filled the missing or unobserved CIFs and standard errors by retaining the estimates from those of the nearest previous observed event time. Next, we transformed the CIFs and standard errors using the complementary log-log transformation and combined them using MIANALYZE procedure (Morisot *et al.*, 2015; Moscovici and Ratitch, 2017). The combined results were again transformed back by the DATA steps. The detailed procedure was illustrated in Figure 1. All analyses were performed using SAS enterprise guide 7.1 (SAS Institute Inc., Cary, NC) in a unix environment. SAS code used for the analyses were provided in an Appendix.

4. Results

Among all the variables of interest in the final cohort, 695 out of 1,371 (50.7%) patients had at least one variable with missing data. The comparison of baseline characteristics between the observed and missing data were summarized in Table 1. Missing data were more likely in those who were diagnosed in the earlier study period (2007-2010), rural residents, those who were not exposed to a statin, those with stage 3 cancer, or those who had fewer medical conditions at diagnosis, indicating that missing may not have occurred completely at random (Stuart *et al.*, 2009).

Among the 1,371 patients in the study cohort, 42 (3.1%) patients experienced an event of hospitalization or ED visit due to HF, 165 (12%) patients died, and 1,164 were censored or remained event-free (no hospital presentation or death) at the end of follow-up period. Given the substantial proportion of mortality observed in the study and that death serves as a competing risk to non-fatal survival outcomes, it would be necessary to conduct competing risk analysis to examine the effect of statin exposure on the HF- related hospitalizations or ED visits.

We examined the quality of the imputed LDL both numerically and graphically. The imputation diagnostic statistics in all 153 imputed datasets were summarized in supplemental Table 1. As shown in the table, none of the absolute difference in means between the observed and imputed ones were greater than two times the standard deviations, nor the variance ratios of the imputed versus the observed values were greater than two or smaller than 0.5. Indeed, the range of the variance ratios was quite narrow from 0.8 to 1.2, further indicating the imputed values did not substantially differ from the observed ones. In addition, we graphically compared the kernel density estimates of the observed to the imputed values with stratification of statin exposure status. The density plot in Figure 2 showed that the distributions of the imputed data were similar to those in subjects with observed values, although some of the imputed values tended to be slightly higher or lower than the observed ones. This subtle discrepancy might reflect uncertainties associated with high proportion of missing values. Nevertheless, both numerical and graphical examinations demonstrated that the imputed and observed data were comparable.

Table 1: Baseline characteristics in complete and missing study data

Variable	Value	Missing	Observed	TOTAL	p-value
		N = 695	N = 676	N = 1,371	
Age	Median (IQR)	71(68 – 75)	71(68 – 74)	71(68 – 75)	0.492
Nearest census based neighbourhood income quintile	Missing	3(0.4%)	0(0.0%)	3(0.2%)	0.208
	1	120(17.3%)	111(16.4%)	231 (16.8%)	
	2	151(21.7%)	132(19.5%)	283(20.6%)	
	3	132(19.0%)	158(23.4%)	290(21.2%)	
	4	129(18.6%)	127(18.8%)	256(18.7%)	
	5	160(23.0%)	148(21.9%)	308(22.5%)	
Rural residence	Missing	1(0.1%)	0(0.0%)	1(0.1%)	0.03
	N	582(83.7%)	598(88.5%)	1,180(86.1%)	
	Y	112(16.1%)	78(11.5%)	190(13.9%)	
Left-sided disease	Missing	3(0.4%)	0(0.0%)	3(0.2%)	0.224
	0	315(45.3%)	303(44.8%)	618(45.1%)	
	1	377(54.2%)	373(55.2%)	750(54.7%)	
LDL level at baseline	Median (IQR)	3 (3-4)	3 (2-3)	3 (2-3)	0.292
Cohort entry year	2007-2009	174(25.0%)	49(7.2%)	223(16.3%)	< .001
	2010-2013	220(31.7%)	246(36.4%)	466(34.0%)	
	2014-2017	301(43.3%)	381(56.4%)	682(49.7%)	
Statin exposure		196(28.2%)	324(47.9%)	520(37.9%)	< .001
Breast cancer stage	1	199(28.6%)	226(33.4%)	425(31.0%)	0.003
	2	315(45.3%)	325(48.1%)	640(46.7%)	
	3	181(26.0%)	125(18.5%)	306(22.3%)	
Hypertension		434(62.4%)	486(71.9%)	920(67.1%)	< .001
Diabetes mellitus		106(15.3%)	202(29.9%)	308(22.5%)	< .001
Chronic obstructive pulmonary disease		121(17.4%)	102(15.1%)	223(16.3%)	0.244
Chronic kidney disease		17(2.4%)	29(4.3%)	46(3.4%)	0.058
Atrial fibrillation		33 (4.7%)	32 (4.7%)	65 (4.7%)	0.99
Myocardial infarction		<= 5(0.7%)	<= 5(0.7%)	10(0.7%)	0.965
Ischemic heart disease without myocardial infarction		55(7.9%)	65 (9.6%)	120 (8.8%)	0.265
Peripheral vascular disease		6(0.9%)	18(2.7%)	24(1.8%)	
Non-statin lipid-lowering drugs		25(3.6%)	49(7.2%)	74(5.4%)	0.003
Angiotensin antagonists		279(40.1%)	345(51.0%)	624(45.5%)	< .001
Beta blockers		133(19.1%)	136(20.1%)	269(19.6%)	0.647
Charlson index	Median (IQR)	0(0 – 6)	0(0 – 6)	0(0 – 6)	0.561
Event status (Hospitalization/ED visit)	0 (alive, no event)	563(81.0%)	601(88.9%)	1,164(84.9%)	< .001
	1 (alive with event)	28(4.0%)	14(2.1%)	42(3.1%)	
	2 (Death)	104(15.0%)	61(9.0%)	165(12.0%)	
Time to event (year)	Median (IQR)	4 (2-7)	4 (2-6)	4 (2-7)	< .001

Subsequently, in each imputed dataset, statin exposed patients were matched with those unexposed at the ratio of 1:1 based on PS derived from the logistic regression model. Although the numbers of matched pairs varied substantially across all the 153 imputed datasets, they appeared to be nearly normal distributed with a median of 259 (IQR 254-264), and minimum and maximum of 237 and 277, respectively. The distribution is shown in Figure 3.

Last, the competing risk analysis was conducted in each PS-matched dataset and then pooled using Rubin's rule. The pooled CIFs of HF-related hospital presentations in the statin exposed and unexposed groups were shown in Figure 4. The CIFs in the top panel showed an abnormal non-monotonic pattern as the plot was solely based on estimates from the observed event times as only

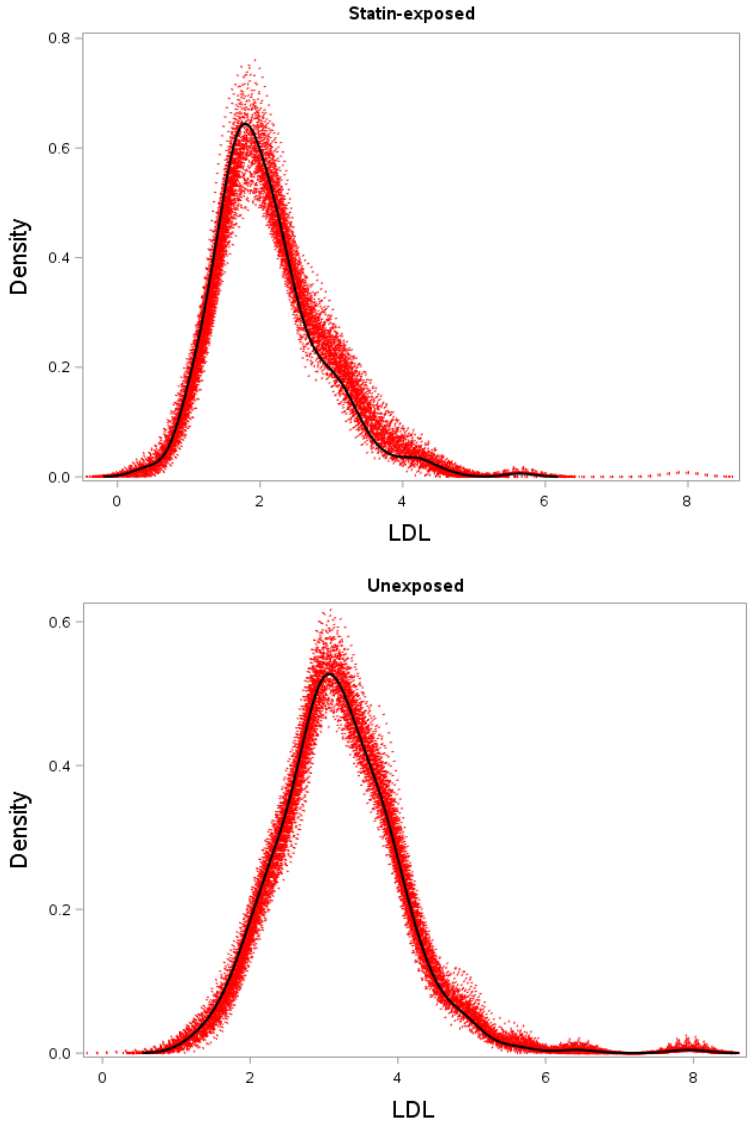


Figure 2: *Probability density function for observed and imputed LDL values stratified by statin exposure status. The solid black line represents the distribution of the observed LDL, while the dashed red lines denote the distribution of the imputed LDL in those subjects with missing LDL. There is one red line for each of the imputed data sets. LDL, low-density lipoprotein.*

these estimates were directly output from the LIFETEST procedure. Since the numbers of matched pairs varied across the 153 imputed datasets, the observed event times differed across all matched datasets. When pooling together, while the observed event times were added up, the corresponding CIFs would be displayed as missing in those datasets without the same observed event times. As

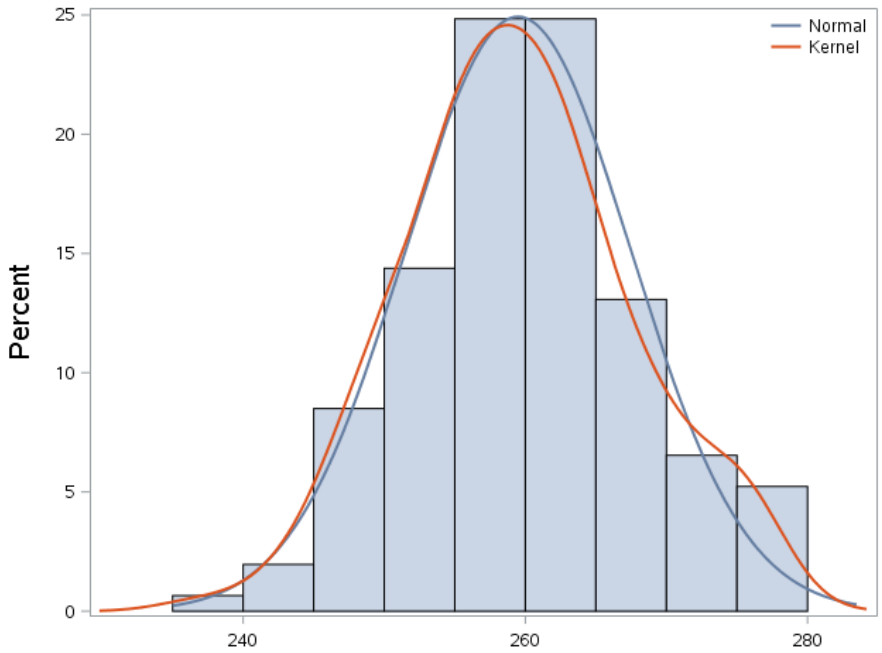


Figure 3: *Distribution of numbers of propensity score-matched pairs in 153 imputed datasets.*

a result, the pooled overall CIFs showed an abnormal incrementation. To correct this, we used the approach of gridded follow-up time in days by computing daily CIFs for the entire follow-up time with missing CIFs and standard errors replaced with values from the nearest previous observed ones in each matched dataset before we pooled them. As such, there was no missing CIF in any of the follow-up time, and the resultant CIF graph in the bottom panel correctly showed a monotonically incremental pattern. The p-value of 0.09 from the Fine-Gray subdistribution hazard model indicated that, the CIFs between the exposed and unexposed groups were not significantly different.

Compared to the statin unexposed patients, the cause-specific hazard ratio for HF-related hospital presentations in the exposed patients was 0.46 (95% 0.20–1.07) with a p-value of 0.07. Therefore, statin exposure decreased the rate of HF-related hospital presentations by 54% in patients who were currently alive and event free, although the difference was not statistically significant at the level of 0.05.

5. Discussion

Observational data with incomplete variable measures are prevalent in biomedical research. MI is a powerful tool to handle missing data with the abilities of increasing statistical power and efficiency (Liu and De, 2015). With the application of PS-matching, observational data can be used to reduce the effects of confounding when estimating the effects of treatment. In this paper, we illustrated explicitly the methodologies of competing risk analysis in incomplete observational data. We used the approaches of MI FCS to impute missing LDL data, then matched statin exposed and unexposed patients by PS and estimated both relative and absolute risk of HF-related hospital presentations. In

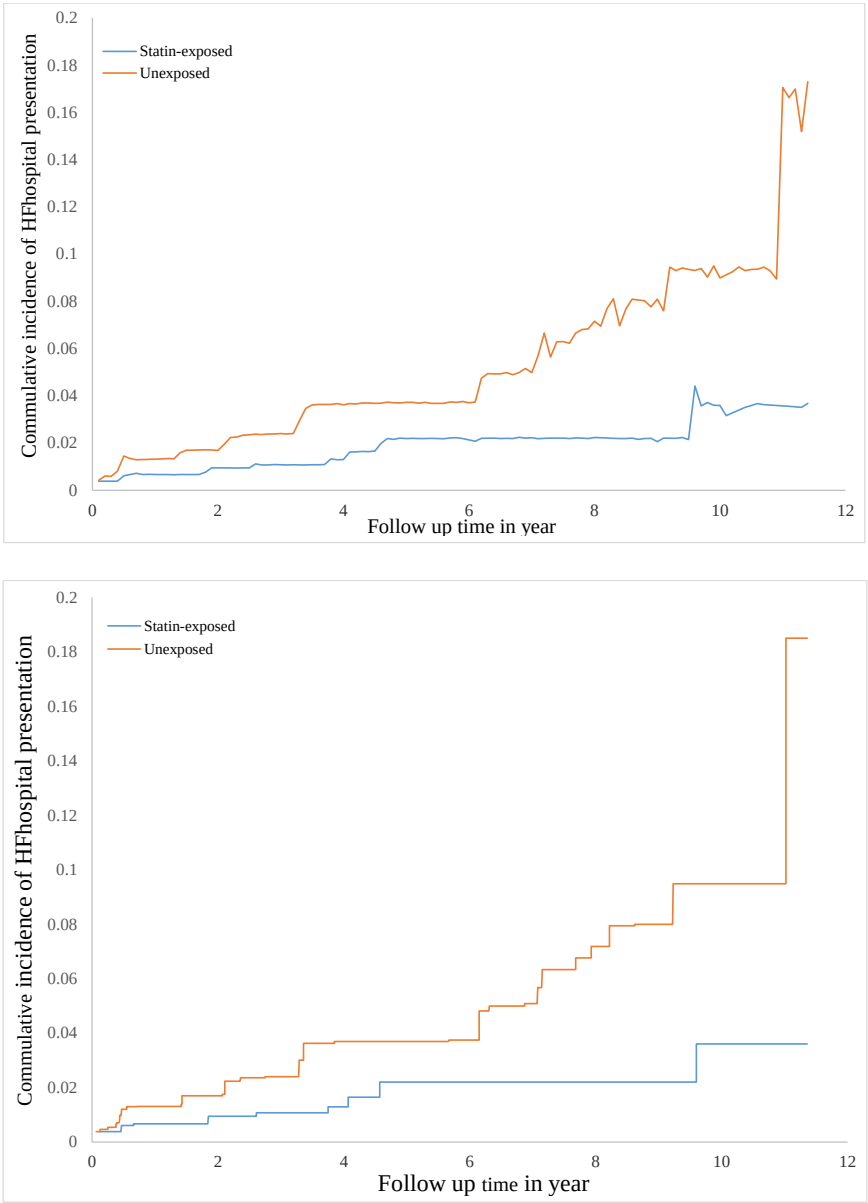


Figure 4: *Overall cumulative incidence functions (CIF) of heart failure-related hospitalization or emergency department visit with stratification of statin-exposure status in propensity score-matched samples. The first panel derived from estimates of observed event times only showed an abnormal patter, and the second panel showed a monotonic incrementing by using estimates from all event time points in day.*

addition, we provided a practical method to generate monotonic CIF graphs in the settings when the observed event times are not consistent across multiple matched samples.

Studies have argued that absolute measures of treatment effect are better than relative measures of treatment effect for clinical decisions making (Jaeschke *et al*., 1995; Laupacis *et al*., 1988). In competing risk analysis, the absolute measures of treatment effect can be estimated through CIFs (Austin and Fine, 2019). Like Kaplan-Meier survival curve, to plot CIFs in a single dataset, usually only estimates from the observed event times are sufficient. However, this may be problematic in multiple imputed and PS-matched samples when individual estimates are pooled and combined due to varying numbers of matched pairs across all matched datasets, which inevitably results in inconsistency of observed event times when pooled together. The approach presented in this paper is easy and straightforward to implement for generating smooth and monotonic CIFs in such settings.

The unbiased estimates with MI depend on correctly specified imputation models for each incomplete variable. FCS MI is an appealing approach in settings when missing data exist in both numeric and categorical variables as separate imputation models can be specified on a variable-by-variable basis (Liu and De, 2015). In addition, an appropriate imputation number is crucial to minimize variability in estimates of regression coefficients, test statistics and p-values across repeated MI analyses. A rule of thumb is that this number should be at least equivalent to the percentages of missing subjects, so that the pooled estimated regression coefficients and standard errors would not vary meaningfully across repeated MI analysis (White *et al*,. 2011; Austin *et al*.,2021). In our study, we initially set the imputation number as 51, which equals to the number of overall missing percentages in the cohort. However, we noticed the pooled p-value for treatment effect fluctuated dramatically in the subsequent competing risk analysis, which implied large variance might be present in these 51 imputed datasets. After we increased the imputation number to 153, three times the number of missing percentages, the direction of p-value was then stabilized consistently.

Like many other survival analyses with incomplete measures of covariates, the assumption that the data were missing at random (MAR) was adopted in this study. Furthermore, we assumed that the missingness is independent of any unobserved information including censoring time. While these assumptions may be plausible in many settings with covariates measured at baseline, uncertainties may arise when the missingness of covariates is associated with future failure time or censoring time. In such cases, more sensible or stringent assumptions may be necessary to take into consideration, such as censoring-ignorable MAR (CIMAR) and failure-ignorable MAR (FIMAR) (Rathouz, 2007). Although it may not be straightforward to implement in practice, further work to optimize MI under these two assumptions for right-censored survival data would be very beneficial.

In this study, we have focused solely on imputation for the missing data in explanatory variables and did not explore MI procedures for missing outcomes in competing risk analysis, such as missing data in the causes of failure. In such cases, different MI models may be employed to accommodate unique nature of these variables. Interested readers may refer to literature for details (Lee *et al*., 2014; Moreno-Betancur and Latouche, 2013; Han *et al*., 2021). Despite this, the same methodologies for subsequent PS-matching and competing risk analysis presented in this paper can still be applied.

We estimated causal treatment effect of statin exposure on HF-related hospital presentation by PS-matching, in which the methodologies have been well established for a treatment variable with two groups. For PS-matching in treatment with more than two groups, the corresponding techniques appear less developed. Nevertheless, a generalized propensity score was proposed to account for multiple levels of treatment (Imbens, 2000), and a three-way matching approach seemed effective with lower or equal bias with little or no cost to mean squared error compared to pairwise or common referent approaches in many study scenarios for three categorical treatments (Rassen *et al*, 2013).

Censoring is very common in survival data and standard methods for survival analysis require that the censoring be noninformative (Allison, 2010). He and colleagues showed that the conventional

Fine-Gray model may result in biased coefficient estimates if the censoring distribution depends on covariates (He *et al.*, 2016). Unfortunately, there is no procedure currently available in SAS to test this assumption or to allow one to incorporate the effect of covariates on the censoring distribution. Consequently, a limitation of the analyses described in the current study is that we were not able to test this assumption of covariate-independent censoring.

In summary, we presented detailed methodologies for conducting competing risk analysis in incomplete observational data with applications of MI and PS-matching, with an emphasis on a practical approach for plotting monotonic CIFs derived from integrated estimates from multiple PS-matched samples.

6. Acknowledgments

This study was supported by the ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care. Parts of this material are based on data and/or information compiled and provided by Canadian Institute of Health Information (CIHI). Parts of this material are based on data and information provided by Cancer Care Ontario. We thank IQVIA Solutions Canada Inc. for use of their Drug Information File. The data sets used for this study are held securely in a linked deidentified form and analyzed at ICES. The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred. Dr. Austin was supported by a Mid-Career Investigator Award from the Heart and Stroke Foundation.

7. Disclosure statement

The authors report no conflicts of interests.

Appendix

SAS codes for multiple imputation, propensity-score matching, competing risk analysis and cumulative incidence function plotting

In this Appendix, we provide SAS codes for multiple imputation, propensity-score matching, competing risk analysis and integration of cumulative incidence function described in the text.

The input data contain one record per subject, the outcome (eventC) for competing risk analysis is defined as one if a patient experienced a hospital presentation due to HF, two if a patient died (competing risk) and zero if no event occurred. The corresponding time to event (timeC) is defined as time in days from index date to a hospital presentation (if event = 1), or death (if event = 2) or being censored at the end of follow-up on December 31, 2018. Statin exposure is coded as dummy variable with one as exposure. Missing values in rural, income quintile, laterality and LDL are left as is without extra coding. X1 represents a list of continuous variables without missing values, X2 represents a list of categorical variables without missing values.

The SAS codes shown below are organized as follows. First, we use PROC MI to impute missing data. Second, in each of imputed datasets, we use Proc Logistic to compute propensity score and then use the macro %gmatch for matching. Last, in each matched dataset, we conduct competing risk analysis and then use Proc MIanalyze to combine and integrate model coefficients and cumulative incidence function.

Step 1, multiple imputation

```
proc MI data = mydata seed = 202207 nimpute = 153 out = MIdata;
```

```
     class eventC rural incquint Laterality X2;
      FCS logistic(rural = timeC eventC X1 X2 incquint Laterality LDL);
      FCS logistic(incquint = timeC eventC X1 X2 rural Laterality  LDL);
      FCS logistic(Laterality = timeC eventC X1 X2 rural incquint LDL);
      FCS regpmm  (LDL = timeC eventC X1 X2 rural incquint Laterality);
      var timeC eventC X1 X2 rural incquint Laterality LDL;
  run;
```

Step 2, propensity-score matching

Only the codes for PS-matching with the first imputed dataset are shown below. PS-matching needs to be done separately in each of 153 imputed datasets.

```
/*Compute propensity score*/
data subcohort;
   set MIdata;
   If _imputation_ = 1;
run;

proc logistic data = subcohort descending;
   class X2 /param = ref ref = first;
   model Statin_Exposure = X1 X2 rural incquint Laterality LDL /lackfit;
   output out=out_ps prob = ps xbeta = logit_ps;
run;

 /*compute standard deviation of the logit of the propensity score*/
 proc means data = out_ps std ;
     var logit_ps;
     output out = out_ps_std (keep = std) std = std;
 run;

 /* Calipers of width = 0.2*standard deviations of the logit of PS*/
 data out_ps_std2 ;
    set out_ps_std ;
    std = 0.2*std;
  run;

  data _null_;
     set out_ps_std2;
     call symput('stdcal', std);
  run;

 /*Match subjects on the logit of the propensity score*/
  data ps_case;
      set out_ps;
      if Statin_Exposure = 1;
```

```
      case_id = _N_;
  run;

  data ps_control;
     set out_ps;
     if Statin_Exposure = 0;
     ctrl_id = _N_;
  run;

proc sort data = out_ps;
     by Statin_Exposure;
run;

data out_ps;
     set out_ps;
     id = _N_;
run;
```

The %gmatch macro from Mayo Clinic Research can be used for PS-matching (https://bioinformaticstools.mayo.edu/research/gmatch/). This macro performs greedy matching. The information of macro parameters can be obtained through the above website. The following codes are based on codes in the chapter 3 – "propensity score matching for estimating treatment effects" by Austin PC *et al* from the book "Analysis of observational health care data using SAS" 2010 (edited by Faries DE *et al*.).

%include 'gmatch.sas';

```
%gmatch(
        Data  = out_ps,
        Group = Statin_Exposure,
          Id  = id,
       mvars  = logit_ps,
         wts  = 1,
         dist = 1,
       dmaxk  = &stdcal,
       ncontls = 1,
       out  = ps_matchpairs,
       seedca  = 202207,
       seedco  = 702022,
        print  = F);

   data ps_matchpairs;
       set ps_matchpairs;
       pair_id = _N_;
   run;

/*Create a dataset containing the matched unexposed patients*/
 data ps_match_ctrl;
```

```
      set ps_matchpairs;
      ctrl_id = __IDC0;
      logit_ps =_C01;
      keep pair_id ctrl_id logit_ps;
 run;

/*Create a dataset containing the matched exposed patients*/
 data ps_match_case;
      set ps_matchpairs;
      case_id = __IDCA;
      logit_ps =_CA1;
      keep pair_id case_id logit_ps;
 run;

 proc sort data = ps_match_ctrl;
      by ctrl_id;
 run;

 proc sort data = ps_match_case;
      by case_id;
  run;

  proc sort data = ps_case;
      by case_id;
  run;

  proc sort data = ps_control;
      by ctrl_id;
  run;

  data ps_match_ctrl;
       merge ps_match_ctrl (in = f1)
                  ps_control;
       by ctrl_id;
       if f1 ;
  run;

  data ps_match_case;
       merge ps_match_case (in = f1)
                  ps_case;
       by case_id;
       if f1;
  run;

  /*Long format by adding together*/
  data ps_match;
```

```
        set ps_match_ctrl
             ps_match_case;
   run;
```

The above matching is done in each of imputed datasets, the corresponding matched datasets are then added together with the variable imputation to distinguish imputation sequence. This new dataset is named as ps_match_all.

Step 3, competing risk analysis

```
/*Hazard ratios by cause-specific competing risk model to estimate treatment
effect in the PS-matched datasets with robust variance estimator to account
for clustering */
ods output ParameterEstimates = Hratio;
proc phreg data = ps_match_all covs(agg);
   class Statin_Exposure/param = ref ref=first;
   model timeC *eventC (0, 2) = Statin_Exposure /rl ties = efron;
   ID pair_id;
   by _Imputation_;
run;
/*Combine estimates from the above model*/
ods output ParameterEstimates = mianal_HR;
proc mianalyze data = Hratio;
   modeleffects estimate;
   stderr StdErr;
run;

/*Exponentiate to compute hazard ratio and confidence interval*/
data mianal_HR;
   set mianal_HR;
   HR = exp(estimate);
   HR_LCL = exp(LCLmean);
   HR_UCL = exp(UCLmean);
   rename Probt = pvalue_comb;
run;

/* p value for CIFs comparison by Sub-distribution competing risk model with
robust variance estimator to account for clustering*/
ods output ParameterEstimates = CIFpvalue;
proc phreg data = ps_match_all covs(agg);
   class Statin_Exposure/param = ref ref = first;
   model timeC *eventC (0) = Statin_Exposure /rl eventcode = 1;
   ID pair_id;
   by _Imputation_;
 run;

/*Combine estimates from the above model*/
ods output ParameterEstimates = mianal_CIFpvalue;
```

```
proc mianalyze data = CIFpvalue;
   modeleffects estimate;
   stderr StdErr;
run;

data mianal_CIFpvalue;
   set mianal_CIFpvalue;
   rename Probt = CIF_pvalue;
run;
```

Step 4, integration of CIFs. First, generate an artificial dataset with three variables including statin exposure, imputation and follow up time in increment of 1 day

```
/*Create a macro variable for the maximum value of the variable timeC*/
proc sql;
   select max(timeC) into: maxfu
   from ps_match_all;
quit;

data days;
   do Days = 0 to &maxfu;
      output;
   end;
run;

data exposure;
   do Statin_exposure = 0 to 1;
      output;
   end;
run;

data impt;
   do _imputation_ = 1 to 153;
      output;
   end;
run;

/*Merge the above three datasets*/
proc sql;
   create table grid as
   select a.*, b.*, c.*
   from exposure as a, impt as b, days as c;
quit;

/*Output observed CIF from Proc Lifetest*/
proc lifetest data = ps_match_all  cs = none notable outcif = CIF ;
   time timeC *eventC (0)/eventcode = 1;
```

```
   strata Statin_Exposure;
   by _Imputation_;
 run;

/*Merge the above two datasets*/
proc sql;
   create table cif_grid as
   select a.*, b.cif,b.stderr, b.timeC
   from  grid as a full join cif as b
   on a._imputation_ = b._imputation_ and a.days = timeC and
      a.Statin_exposure = b.Statin_exposure
   order by Statin_exposure,_imputation, days;
quit;

/*Fill out missing values of CIF by carrying over the previous non-missing values*/
data cif_grid;
   set cif_grid;
if days = 0 then do;
       cif = 0;
       stderr = 0;
   end;

   retain CIF_N;
   retain stderr_N;
   if not missing(cif)    then cif_N   = cif;
   if not missing(stderr) then stderr_N = stderr;
run;

/*Transform the CIF and SE*/
data CIF_1 CIF_2;
   set CIF_grid;
   if 0 < CIF_N < 1 then do;
      CIF_tm = log(-log(CIF_N));
      stderr_tm = sqrt((1/(log(CIF_N))**2)*((StdErr_N**2)/(CIF_N**2)));
      output CIF_1;
   end;
   else output CIF_2; /*for CIF = 0 or 1 estimates*/
run;

/*Combine estimates from all imputed datasets using MIANALYZE*/
proc sort data = CIF_1;
   by Statin_Exposure days _Imputation_;
run;

ods output ParameterEstimates = mianalCIF;
proc mianalyze data = CIF_1;
```

```
   modeleffects CIF_tm;
   stderr StdErr_tm;
   by Statin_Exposure days;
run;

/*Back-transform the combined CIF estimates and compute CI. The dataset
mianalCIF2 can be used to plot integrated CIF*/
Data mianalCIF2;
   set mianalCIF;
   format CIF_comb CIF_StdErr_comb CIF_LCL_comb CIF_UCL_comb zRight
          8.6;
   CIF_comb = exp(-exp(Estimate));
   CIF_StdErr_comb = abs(CIF_comb * StdErr * log(CIF_comb));
   zRight = quantile("Normal", 1-0.05 / 2);
   CIF_LCL_comb = CIF_comb**(exp(zRight*StdErr));
   if CIF_LCL_comb<0 then CIF_LCL_comb = 0;
   CIF_UCL_comb = CIF_comb**(exp(-zRight*StdErr));
run;
```

References

Abdel-Qadir H, Bobrowski D, Zhou L, Austin PC, Calvillo-Argüelles O, Amir E, Lee DS, and Thavendiranathan P (2021). Statin exposure and risk of heart failure after anthracycline-or trastuzumab-based chemotherapy for early breast cancer: A propensity score– matched cohort study, *Journal of the American Heart Association*, **10**, e018393.

Allison PD (2010). *Survival Analysis Using SAS: A Practical Guide (2nd ed)*, SAS Press, Cary, NC. SAS Institute Inc.

Austin PC (2011). An Introduction to propensity score methods for reducing the effects of confounding in observational studies, *Multivariate Behavioral Research*, **46**, 399–424.

Austin PC, Lee DS, and Fine JP (2016). Introduction to the analysis of survival data in the presence of competing risks, *Circulation*, **133**, 601–609.

Austin PC and Fine JP (2019). Propensity-score matching with competing risks in survival analysis, *Statistics in Medicine*, **38**, 751–777.

Austin PC, White IR, Lee DS, Buuren SV, Buuren LV, and Methodology and Statistics for the Behavioural and Social Sciences (2021). Missing data in clinical research: A tutorial on multiple imputation, *Canadian Journal of Cardiology*, **37**, 1322–1331.

Boyko EJ (2013). Observational research-opportunities and limitations, *Journal of Diabetes and Its Complications*, **27**, 642–648.

Han S, Tsui KW, Zhang H, Kim GA, Lim YS, and Andrei AC (2021). Multiple imputation analysis for propensity score matching with missing causes of failure: An application tohepatocellular carcinoma data, *Statistical Methods in Medical Research*, **30**, 2313–2328.

He P, Eriksson F, Scheike TH, and Zhang MJ (2016). A proportional hazards regression model for the sub-distribution with covariates adjusted censoring weight for competing risks data, *Scandinavian Journal of Statistics*, **43**, 103–122.

Huque MH, Carlin JB, Simpson JA, and Lee KJ (2018). A comparison of multiple imputation methods for missing data in longitudinal studies, *BMC Medical Research Methodology*, **18**, 168–184.

Imbens G (2000). The role of the propensity score in estimating dose-response functions, *Biometrika*, **87**, 706–710.

Jaeschke R, Guyatt G, Shannon H, Walter S, Cook D, and Heddle N (1995). Basic statistics for clinicians: 3.assessing the effects of treatment: Measures of association, *Canadian Medical Association Journal*, **152**, 351–357.

Laupacis A, Sackett DL, and Roberts RS (1988). An assessment of clinically useful measures of the consequences of treatment, *New England Journal of Medicine*, **318**, 1728–1733.

Lee M, Dignam J, and Han J (2014). Multiple imputation methods for nonparametric inference on cumulative incidence with missing cause of failure, *Statistics in Medicine*, **33**, 4605–4626.

Little RJ and Rubin DB (1987). *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

Liu Y and De A (2015). Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study, *International Journal of Statistics in Medical Research*, **4**, 287–295.

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.

Morisot A, Bessaoud F, Landais P, Rébillard X, Trétarre B, and Daurès JP (2015). Prostate cancer: Net survival and cause-specific survival rates after multiple imputation, *BMC Medical Research Methodology*, **15**, 54–68.

Moscovici JL and Ratitch B (2017). Combining Survival Analysis Results after Multiple Imputation of Censored Event Times, In *Proceedings of PharmaSUG 2017 - Paper SP05*, add city.

Nguyen CD, Carlin JB, and Lee KJ (2017). Model checking in multiple imputation: An overview and case study, *Emerging Themes in Epidemiology*, **14**, 8–20.

Rathouz PJ (2007). Identifiability assumptions for missing covariate data in failure time regression models, *Biostatistics*, **8**, 345–356.

Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, and Schneeweiss S (2013). Matching by propensity score in cohort studies with three treatment groups, *Epidemiology*, **24**, 401–409.

Stuart EA, Azur M, Frangakis C, and Leaf P(2009). Multiple imputation with large data sets: A case study of the children's mental health initiative, *American Journal of Epidemiology*, **169**, 1133–1139.

White IR, Royston P, and Wood AM (2011). Multiple imputation using chained equations: Issues and guidance for practice, *Statistics in Medicine*, **30**, 377–399.

Received December 15, 2021; Revised September 20, 2022; Accepted September 22, 2022