Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Imputing missing covariate values in the presence of competing risks

Due to its flexibility, practicability and efficiency, multiple imputation by chained equations is widely used to impute missing data. To avoid biases in the substantive model, it is well known that the imputation model must include all the variables from the analysis model, including the outcome. In 2009, White and Royston showed that when the Cox model is used for the analysis of survival data, the imputation model for each covariate should include the event indicator and the cumulative baseline hazard estimated by the Nelson–Aalen estimator (White IR, Royston P, Stat Med. 2009).

In the competing risks setting, multiple imputation has been proposed only to impute missing information on the cause of failure, and has mostly been used in the analysis of cumulative incidence functions. We extend the work of White and Royston to impute missing covariates in a competing risks setting, where the substantive model is either a cause–specific proportional hazards model or a sub–distribution proportional hazards model. We show that the event indicator and the cumulative baseline hazard of all the competing events should be included in the imputation model. Consequently, this suggests that even in the standard survival analysis framework, the cumulative baseline hazard of censoring should be included in the imputation model.

These approaches are evaluated by a simulation study, and then applied to a real sample of 278 adult patients with acute myeloid leukaemia. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: class file; LaTeX 2_ε; *Statist. Med.*

1. Introduction

All biostatisticians have been faced with missing data, even in the analysis of randomized clinical trial data. Indeed, while much effort is done to avoid missing outcomes, that is, patients lost–to–follow–up, less attention is often made on the record of covariates besides inclusion criteria or stratification variables that are only hardly monitored. Moreover, some covariates may be missing because of the lack of available measures due to invasive or costly medical procedures leading to high missing rates (possibly above 30%).

Popular approaches for analyzing partially observed datasets include complete case analysis which discards the incomplete cases, direct likelihood methods, inverse probability weighting, and multiple imputation (MI) ([**?**]). The latter approach, which has become increasingly attractive for researchers, is a Monte-Carlo method in which the missing values are replaced by m suitably imputed values. From a bayesian perspective, it consists in drawing plausible values of each missing data in a conditional distribution of missing data given the observed ones. This *posterior* distribution represents the information about the missing values that is contained in the observed data ([**?**]).Each of the m imputed dataset is then analyzed independently to obtain a set of m estimates. Final estimates are averaged across the m estimates and standard errors are computed according to the Rubin's rule ([**?**, **?**]).

Due to its efficiency multiple imputation is widely used to impute missing data using either joint multinormal distributions or chained equations. Indeed, missing data are usually multivariate since they concern more than one variable in the sample; the resulting multivariate distribution of the missing data could be generated jointly by numerical techniques throughout Markov chain Monte Carlo algorithm (MCMC) and joint multinormal distributions ([**?**, **?**, **?**]). Nevertheless, to allow more flexibility, in particular when datasets present a mixture of continuous and discrete variables, the method of multiple imputation by chained equations (MICE) also known as fully conditional specification (FCS) is particularly

[†] Please ensure that you use the most up to date class file, available from the SIM Home Page at
www.interscience.wiley.com/jpages/0277-6715

attractive ([**?**, **?**]). It consists of separating the multivariate imputation model into a sequence of univariate imputation models and solve the multivariate case by iteration. In practice, MICE specifies the multivariate imputation model on a variable-by-variable basis by a series of conditional densities, one for each incomplete variable. For each incomplete variable, the user specifies a conditional distribution for the missing data given the other data. Under the assumption that a multivariate distribution exists from which these conditional distributions can be derived, MICE could construct a Gibbs sampler from the specified conditionals. Although, its statistical properties are less well understood than those of the multinormal MI method, simulation studies and recent theoretical results show its great efficiency and robustness in various settings. ([**?**, **?**]).

All those approaches are considered to be valid under Missing at Random (MAR) assumption, *i.e.* if the probability of data being missing does not depend on the unobserved data, conditionally on the observed data. Nevertheless, this validity is also determined by the specification of the imputation model. One way to well specify the imputation model is to define a model congenial or compatible to the analysis model meaning that both models should be compatible with some larger model for the data. Thus, the choice of the analysis model has implications on the definition of the imputation model and especially on the variables which should be included in it. This is one of the reasons why one should consider all the covariates of the analysis model and the outcome in the imputation model ([**?**]).

In the particular setting of survival data, the outcome is defined by a binary event indicator $D \in \{0, 1\}$ where $D = 1$ denotes the occurrence of the event and $D = 0$ the right censoring, and a time of event or censoring, respectively, T. Let suppose a standard Cox proportional hazard model regression model involving an incomplete covariate. As imputation model should involve the outcome as predictor, D and T should be included in the model. It is generally obtained by direct inclusion of D and T (or $\log(T)$) in the imputation model ([**?**, **?**]).

Nevertheless, estimations may still be biased, even using a MICE procedure with predictive mean matching as recommended by Marshall ([**?**]). In 2009, I. White and P. Royston showed that the imputation model should include the event indicator and the cumulative baseline hazard instead of T (or $\log(T)$), and therefore recommended to include the Nelson-Aalen estimator of the cumulative hazard in the imputation model ([**?**]).

In the competing-risks setting, subjects may experiment one out of K distinct and exclusive events. Outcome thus is defined by an event indicator $\varepsilon \in \{0, 1, \ldots, K\}$ where $\varepsilon = 1, \ldots, K$ denotes the observed type of failure and $\varepsilon = 0$ denotes the right censoring and the observed event or censoring time T. Two main approaches have been proposed. The most common approach models the cause-specific hazard of the event of interest while the second approach models the subdistribution hazard associated to the cumulative incidence function ([**?**, **?**]). If the first quantity is easily handle as the "instantaneous risk" per time unit of failure at time t from event of interest given survival till just before t, it has no direct correspondence with the cumulative incidence. On the opposite, the subdistribution hazard is direclty linked to the cumulative inference but has no standard epidemiological sense and then difficult to interpret ([**?**]). Thus, the choice between both approaches depends on the type of type of competing risks and the aim of the study (focused on the cumulative incidence or not).

In presence of missing data in the specific setting of competing risks, previously published works have considered multiple imputation strategies when the cause of failure is missing for some patients, when modelling either one of these quantities: the cause-specific hazard function ([**?**, **?**]), the cumulative incidence function directly ([**?**, **?**]) or through the Fine and Gray model ([**?**, **?**]).

As far as we know in this competing risks setting, only one recent paper focused on missing covariate values and did not used multiple imputation but a bivariate model for both the missing covariates and the missing data mechanism ([**?**]).

We aimed to extend the work of I. White and P. Royston [**?**] to the competing-risks setting by including in the imputation model the cumulative hazard associated with the hazard functions of the analysis model. Notably, we will show that cumulative hazards of all the competing events should be included in case of cause–specific analysis.

In Section 2, we present the notation and the competing risks framework. In Section 3, we state explicit formulae of an imputation model for competing risks data analyzed through a cause–specific hazards Cox model. Section 4 provides simulation results to show the relevance of the approach in finite samples. In Section 5, we illustrate the results using data from a clinical trial in de novo acute myeloid leukaemia (AML). In Section 6, we give a brief discussion.

2. Notations and Assumptions

Suppose a competing-risks setting, in which subjects may fail from one out of K distinct and exclusive causes of failure. Let (T, ε), denote the survival data where T is the minimum of the failure time T^* and the right-censoring time C, $\varepsilon \in \{1, ..., K\}$ denote the failure cause and $\varepsilon = 0$ denotes a right-censored observation. A set of complete covariates Z and X a single incomplete covariate are also defined. Moreover, we assume a non informative censoring independent from Z and X.

We assume that missing value of X may be due to either a missing completely at random (MCAR, probability of missing data is independent of the observed and unobserved data) or a missing at random (MAR) mechanisms.

$\forall k \in \{1, ..., K\}$, the cumulative incidence of the failure from cause k (also known as sub–distribution) is defined by $F_k(t|X,Z) = P(T \leq t, \varepsilon = k|X,Z)$ and the global survival function by $S(t) = 1 - \sum_{k=1}^{K} F_k(t)$.

To rely the hazard of failure from cause k to the baseline recorded covariates X and Z, the functions of interest are either :

(i) the cause–specific hazard (CSH) for this failure cause $\varepsilon = k$,

$$h_k(t|X,Z) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, \varepsilon = k|T \geq t, X, Z)}{\delta t} = \frac{1}{S(t)} \frac{\delta F_k(t)}{\delta t}$$

and its corresponding cumulative hazard $H_k(t)$,

or (ii) the sub–distribution hazard associated with the cumulative incidence of this failure cause,

$$\lambda_k(t|X,Z) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t, \varepsilon = k|T \geq t \cup (T \leq \cap \varepsilon \neq k), X, Z)}{\delta t} = -\frac{\delta \log(1 - F_k(t|X,Z))}{\delta t}$$

and its corresponding cumulative hazard $\Lambda_k(t)$.

3. Imputation model for the Cause-specific approach

We assume that a cause–specific proportional hazards model is valid for $h_k(t)$ for each cause of failure $\varepsilon = k$:

$$h_k(t|X,Z) = h_{k0}(t) \exp(\beta_{kX} X + \beta_{kZ} Z)$$

where β_{kX} and β_{kZ} are unknown regression coefficients to be estimated and $h_{k0}(t)$ are the cause–specific baseline hazards for the cause of failure k.

Our aim is to derived an imputation model for X compatible with this analysis model for $h_k(t)$. The log–likelihood function of the observation (T, ε, X, Z), is given by:

$$\begin{aligned}
\log(P(T,\varepsilon|X,Z)) &= \sum_{k=1}^{K} \left\{ 1_{[\varepsilon=k]} \log(h_k(T|X,Z)) \right\} + \log(S(T|X,Z)) \\
&= \sum_{k=1}^{K} \{ 1_{[\varepsilon=k]} \log(h_k(T|X,Z)) - H_k(T|X,Z) \} \\
&= \sum_{k=1}^{K} \{ 1_{[\varepsilon=k]} \log(h_{k0})(T) + 1_{[\varepsilon=k]}(\beta_{kX} X + \beta_{kZ} Z) - H_{k0}(T) \exp(\beta_{kX} X + \beta_{kZ} Z) \}
\end{aligned}$$

where $H_{k0}(T) = \int_0^T h_{k0}(t) dt$ is the cumulative cause–specific baseline hazard of failure from cause k.

The Bayes theorem is applied to obtain the conditional distribution of X given the observed covariates:

$$\log(P(X|T,\varepsilon,Z)) = \sum_{k=1}^{K} \left\{ 1_{[\varepsilon=k]}(\beta_{kX} X + \beta_{kZ} Z) - H_{k0}(t) \exp(\beta_{kX} X + \beta_{kZ} Z) \right\} + \log(P(X|Z)) + C \qquad (1)$$

where the constant C may depend on ε, T and Z, but not on X.

Let suppose, first, that X is Bernoulli distributed depending on Z through a logistic regression model logit$(p(X = 1|Z)) = \zeta_0 + \zeta_1 Z$. We obtain then

$$\text{logit}(p(X = 1|T,\varepsilon,Z)) = \zeta_0 + \zeta_1 Z + \sum_{k=1}^{K} \left\{ 1_{[\varepsilon=k]} \beta_{kX} \right\} - \sum_{k=1}^{K} \{ H_{k0}(t) \exp(\beta_{kZ} Z)(\exp(\beta_{kX}) - 1) \} \qquad (2)$$

In absence of Z, the equation 2 reduces itself to

$$\zeta_0 + \sum_{k=1}^{K} \left\{ 1_{[\varepsilon=k]} \beta_{kX} \right\} + \sum_{k=1}^{K} \{ (\exp(\beta_{kX}) - 1) H_{k0}(T) \}$$

i.e., depends only on the cause of failure ε and the cumulative hazard $H_{k0}(T)$ of all causes of failure.

Following the development proposed by White and Royston in their appendix in presence of Z, a regression model for X could be approximated by a logistic model depending on ε, Z, and $H_{k0}(T)$ for all k ([**?**]):

$$\text{logit}(p(X=1|T,\varepsilon,Z)) \approx \zeta_0' + \zeta_1 Z + \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\beta_{kX}\right\} + \sum_{k=1}^{K}\{\beta_{kX}' H_{k0}(t)\} \tag{3}$$

This approximation using Taylor series approximation is valid when all β_{kZ} have small variance. Moreover, a Taylor expansion of higher degree shows that interactions terms $H_{k0}(T)Z$ for all k improve the approximation.

If X is normally distributed, the relationship given (T,ε,Z) is not truly linear on ε, Z, and $H_{k0}(T)$ for all k. Nevertheless, as shown by White and Royston for small var$(\beta_{kX}X+\beta_{kZ}Z)$, a good approximation for $(X|T,\varepsilon,Z)$ could be a linear regression including ε, Z, and $H_{k0}(T)$ as predictors. As for binary X, a more accurate approximation assuming small var$\beta_{kX}X$ and small $H_{k0}(T)$ show that the linear model could be improved by including also interaction terms $H_{k0}(T)Z$ for all k.

This result suggests that the bazeline cumulative cause–specific hazard H_{k0} of all the causes of failure k should be included in the imputation model. Thus, our hypothesis was that logistic or linear regression model of X on ε, Z and the bazeline cumulative cause-specific hazard H_{k0} of all the causes of failure k may provide good imputation model for X in case of cause–specific analysis. We propose to approximate the baseline cumulative cause–specific hazard by the Nelson-Aalen estimator of the cumulative cause–specific hazard of all the causes of failure ([**?**, **?**]).

4. Simulation study

Suppose two competing events, the event 1 being the event of interest. Let be t_1 and t_2 the latent failure times of cause 1 and 2, respectively. No censoring were considered.

4.1. Data generating process

Covariates X and Z were generated using a multinormal distribution with means 1, variances 1 and a correlation of 0. Two missing mechanisms were evaluated (i) a MCAR mechanism with probability of 0.5 of missingness for X and (ii) a MAR mechanism defined by a probability of missingness for X of 0.1 if $Z>0$ and of 0.9 if $Z \leq 0$.

Times of failure from cause 1 (t_1) were generated according to a Weibull distribution with shape a_1 and scale $h_{10} \times exp(-(\beta_{1X}X+\beta_{1Z}Z)/a_1)$. Times to failure from cause 2 (t_2) were generated either from a Weibull distribution depending on X and Z with shape=a_2 and scale=$h_{20} \times exp(-(\beta_{2X}X+\beta_{2Z}Z)/a_2)$, or a uniform distribution on $[0;0.5]$ independently from X and Z.Parameter values were fixed to obtain about 50% of type 1 events at $a_1=0.3$, $h_{10}=1$, $\beta_{1X}=0.5$, $\beta_{1Z}=0.5$, $a_2=1.7$, $h_{20}=0.5$, $\beta_{2X}=-0.5$, $\beta_{2Z}=0.5$. The time to failure t was defined as the minimum of t_1 and t_2. Percentages of event 1 obtained were 53% and 49% with the Weibull distribution and the uniform distribution for t_2 respectively.

The total number of patients was fixed at 200. A total of $N=1000$ independent datasets were generated for each setting.

4.2. Analysis Methods

The following proportional hazards model was the analysis model

$$h_1(t|X,Z) = h_{10}(t)\exp(\beta_{1X}X+\beta_{1Z}Z).$$

First, data were analysed before any data deletion as a benchmark for the MI procedure (*Ref*). Secondly, missing data were generated as described above. A Complete Case Analysis (*CCA*) using only the complete observations was performed. Then, incomplete data were imputed $m=10$ times using two different imputation models. The first imputation model for X was a linear regression model including Z, ε and only the cumulative hazard of event 1, $\hat{H}_1(T)$ estimated by Nelson-Aalen estimator, by analogy to the imputation models proposed in standard survival analyses with a Cox model; this will be further denoted as CH_1. The second imputation model for X was a linear regression model including Z, ε and both the cumulative hazards of event 1, $\hat{H}_1(T)$ and of event 2, $\hat{H}_2(T)$ estimated by Nelson-Aalen estimator (CH_{12}). The imputed data sets were analysed by fitting analysis models and estimates were pooled using Rubin's rules ([**?**]).

4.3. Implementation

Simulation study was conducted under the statistical R package ([**?**]). The Nelson-Aalen estimator was provided by the *survival* R package ([**?**]). The overall imputation procedure was implemented using the *mice* R package ([**?**]). Imputations

Table 1. Simulation results for β_{1X}. MAR missing mechanism for X. risk prop

	MAR				
	$\beta_{1X} = 0$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	-0.001	0.140	0.143	0.946	0.054
CCA	0.001	0.242	0.251	0.945	0.055
CH12	0.042	0.214	0.209	0.951	0.049
CH12Z	-0.019	0.224	0.220	0.947	0.053
CH1	0.033	0.224	0.213	0.961	0.039
CH1Z	-0.040	0.239	0.228	0.023	0.052
	$\beta_{1X} = 0.5$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	0.506	0.149	0.151	0.948	0.931
CCA	0.519	0.259	0.269	0.945	0.535
CH12	0.492	0.233	0.229	0.949	0.555
CH12Z	0.431	0.244	0.236	0.940	0.425
CH1	0.493	0.244	0.234	0.950	0.520
CH1Z	0.438	0.260	0.251	0.427	0.392
	$\beta_{1X} = 1$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	1.014	0.157	0.157	0.953	1.000
CCA	1.035	0.273	0.283	0.946	0.977
CH12	0.926	0.236	0.223	0.948	0.975
CH12Z	0.859	0.249	0.230	0.926	0.925
CH1	0.931	0.237	0.225	0.951	0.972
CH1Z	0.887	0.252	0.235	0.945	0.926

were performed using the approximate proper imputation algorithm ([**?**]). Rubin Rules were applied using the Barnard-Rubin adjusted degrees of freedom for potential small samples ([**?**]).

The performance of each method was assessed by computing the empirical mean of the parameter estimates ($Mean$), the relative bias ($Rbias$), the root mean square of estimated standard errors (Se_{Cal}), the empirical Monte Carlo standard errors (Se_{Emp}) and the coverage of nominal 95% confidence intervals (95%CI) for β_{1X} and β_{1Z} ($Cover.$). Note that when observed coverage is 95% from 1000 simulations, a 95%CI for true coverage is [93.6–96.4%].

4.4. Results

Results of the simulation study are given in Table 1 and Table 2. As expected, in absence of any missing data, observed biases are small (always below 2.2%), standard errors are minimal and estimated coverages are around 95%.

Concerning $\hat{\beta}_{1X}$. Point estimates obtained with CCA analysis are always slightly above 0.5 with a bias systematically above 2%. With the CH_1 imputation approach, results seem always negatively biased whatever the missing data mechanism and the shape of t_2 distribution, with relative biases ranging from -26% to -7.4%. The coverage is then systematically below 95%. With CH_{12}, the point estimates are closer to 0.5 whatever the missing mechanism and the shape of t_2 distribution. Empirical and calculated standard errors are similar and close to the observed value with CCA. Coverage is close to 95% except for the MAR mechanism and uniform distribution for t_2.

Concerning $\hat{\beta}_{1Z}$. Point estimates obtained with CCA analysis are still slightly above 0.5 with a bias systematically above 2%. With CH_1, pointes estimates are close to 0.5 with relative biases ranging from -2.4% to 6.6%. Empirical and calculated standard errors are similar and coverage is close to 95%. With CH_{12}, point estimates are close to 0.5 with relative biases ranging from -0.4% to 2.2%. Empirical and calculated standard errors are similar and smaller than those observed with CCA. Coverage is close to 95%, ranging from 93.7% to 95.9%.

Otherwise, simulations studies showed that estimated $\hat{\beta}_{1X}$ are less biased with CH_{12} than with CH_1 and that precision for $\hat{\beta}_{1Z}$ is better with CH_{12} than with CCA.

Table 2. Simulation results for β_{1X}. MAR missing mechanism for X. prop haz

	MAR				
	$\beta_{1X} = 0$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	-0.004	0.163	0.167	0.947	0.053
CCA	-0.009	0.293	0.307	0.948	0.052
CH12	0.012	0.253	0.218	0.975	0.025
CH12Z	-0.015	0.269	0.245	0.958	0.042
CH1	-0.032	0.248	0.198	0.984	0.016
CH1Z	-0.065	0.259	0.217	0.025	0.029
	$\beta_{1X} = 0.5$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	0.512	0.179	0.181	0.953	0.833
CCA	0.534	0.328	0.345	0.945	0.390
CH12	0.393	0.277	0.236	0.955	0.276
CH12Z	0.406	0.296	0.264	0.953	0.297
CH1	0.325	0.276	0.234	0.924	0.187
CH1Z	0.317	0.292	0.253	0.335	0.178
	$\beta_{1X} = 1$				
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$
Ref	1.027	0.208	0.211	0.950	0.999
CCA	1.075	0.391	0.414	0.951	0.841
CH12	0.732	0.309	0.246	0.899	0.700
CH12Z	0.768	0.326	0.262	0.931	0.687
CH1	0.683	0.312	0.258	0.850	0.620
CH1Z	0.683	0.329	0.269	0.870	0.572

Table 3. Simulation results for β_{1X}. MAR missing mechanism for X. non prop haz

	MAR					
	$\beta_{1X} = 0$					
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$	
ref.1	0.001	0.108	0.108	0.952	0.048	0.001
cc.1	0.001	0.173	0.176	0.947	0.053	0.001
ch12	0.008	0.176	0.166	0.960	0.040	0.008
ch12z	0.013	0.179	0.172	0.952	0.048	0.013
ch1	-0.008	0.176	0.160	0.965	0.035	-0.008
ch1z	0.005	0.183	0.171	0.001	0.040	-0.995
	$\beta_{1X} = 0.5$					
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$	
Ref	0.506	0.149	0.151	0.948	0.931	
CCA	0.519	0.259	0.269	0.945	0.535	
CH12	0.492	0.233	0.229	0.949	0.555	
CH12Z	0.431	0.244	0.236	0.940	0.425	
CH1	0.493	0.244	0.234	0.950	0.520	
CH1Z	0.438	0.260	0.251	0.427	0.392	
	$\beta_{1X} = 1$					
	$Mean$	Se_{cal}	Se_{emp}	$Cover.$	$power$	
Ref	1.014	0.157	0.157	0.953	1.000	
CCA	1.035	0.273	0.283	0.946	0.977	
CH12	0.926	0.236	0.223	0.948	0.975	
CH12Z	0.859	0.249	0.230	0.926	0.925	
CH1	0.931	0.237	0.225	0.951	0.972	
CH1Z	0.887	0.252	0.235	0.945	0.926	

Table 4. Coefficients, standard errors and p.values associated to the risk of relapse according to the methods.

	Complete Case		Death Cum. hazard		Relapse and Death Cum. hazard	
	Coef (SE)	p.value	Coef (SE)	p.value	Coef (SE)	p.value
Arm	-0.400 (0.256)	0.120	-0.329 (0.185)	0.076	-0.323 (0.186)	0.082
Age	0.333 (0.260)	0.200	0.418 (0.197)	0.034	0.429 (0.199)	0.031
CD33	0.247 (0.672)	0.710	0.122 (0.552)	0.826	0.129 (0.526)	0.808
NMT3	0.515 (0.268)	0.055	0.488 (0.270)	0.073	0.525 (0.265)	0.048

Table 5. Coefficients, standard errors and p.values associated to the risk of death without relapse according to the methods.

	Complete Case		Relapse Cum. hazard		Relapse and Death Cum. hazard	
	Coef (SE)	p.value	Coef (SE)	p.value	Coef (SE)	p.value
Arm	-0.344 (0.464)	0.460	-0.329 (0.247)	0.185	-0.321 (0.242)	0.184
Age	0.042 (0.462)	0.930	0.40 (0.258)	0.120	0.402 (0.257)	0.117
CD33	0.209 (1.27)	0.870	0.734 (0.720)	0.308	0.706 (0.721)	0.326
NMT3	0.466 (0.486)	0.340	0.381 (0.451)	0.400	0.433 (0.371)	0.245

5. Motivating Example

We applied our work on data from a randomized clinical trial (ALFA-0701) conducted in 278 patients with de-novo acute myeloid leukaemia (AML) aiming to evaluate the benefit of gemtuzumab ozogamicin (GO), an anti-CD33 antibody conjugate, to the standard treatment on the event-free survival (EFS) ([**?**]). Such an endpoint is actually a composite endpoint defined by the time to either relapse or death free of relapse. We were interested in further assessing the effect of the treatment on the two components of the end point, namely time to relapse (132 events, 47%) and time to death without relapse (78 events, 28%), adjusting on three prognostic factors, namely CD33 blast cells count (considered as a continuous variable after log-transformation), DNMT3 mutation (considered as a binary variable) and age above below 60 years. This defines a competing risks framework (Figure **??**).

Among these 278 included patients, 37 (13%) had missing data on CD33 blast cells count, while the missing reaches 131 patients (47%) on DNMT3 mutation. Corresponding complete case analysis results in 130 patients, that is 47% of the whole sample, with 68 (52%) patients who relapsed and 20 (15%) who died free of relapse.

Table 4 and Table 5 display the estimated effects of the GO arm adjusted on age, CD33 blast cells and DNMT3 mutation, based on CCA and MICE with $m = 50$, 10 iterations. Imputation models were a linear regression and a logistic regression model for CD33 blast cells and DNMT3 mutation, respectively. The analysis model was a proportional cause–specific hazard model for the hazard of relapse or the hazard of death free of relapse. Three approaches were considered, namely; (i) Complete Case Analysis, (ii) Multiple imputation considering only the cumulative hazard of interest, (iii) Multiple imputation considering both cumulative hazards.

Not surprisingly, those results confirm the interest of multiple imputation over the complete case analysis. Indeed, precision on Arm or Age effect is improved with both multiple imputation approaches (Table 4 and Table 5). The impact of adding the cumulative hazards of all causes of failure was light. Nevertheless, the estimated regression coefficient obtained for $NMT3$ (Table 4 and Table 5) increases similarly to those observed in the simulation study. Moreover, the impact of $NMT3$ mutation on the hazard of relapse becomes significant (Table 4). The impact on the $CD33$ coefficients was almost null, probably due to a smallest percentage of missing data for $CD33$ blast cells than for $DNMT3$ mutation.

6. Further development for cumulative incidence hazard

In the presence of competing risks, one could be interested in the sub–distribution hazard (*i.e.*, the hazard directly associated with the cumulative incidence function) that could be modelled via a proportional hazards model known as the Fine and Gray model ([**?**]). It has been shown that both approaches (cause–specific hazard and sub–distribution hazard models) could be modelled via proportional hazards assumptions but are not compatible ([**?**]). Thus, the use of a Fine and Gray model as the analysis model should modify our way to define the imputation model at least on a congeniality argument. Then, we propose also an imputation model in this setting.

We assume that a subdistribution proportional hazard model is valid for $\lambda_k(t)$ for each cause of failure k:

$$\lambda_k(t|X,Z) = \lambda_{k0}(t)\exp(\beta_{kX}X + \beta_{kZ}Z)$$

where β_{kX} and β_{kZ} are unknown regression coefficients to be estimated and $\lambda_{k0}(t)$ the baseline subdistribution hazard for the cause of failure k. The likelihood function of the observations (T,ε,X,Z), is given by:

$$\begin{aligned}
P(T,\varepsilon|X,Z) &= \prod_{k=1}^{K}\left\{\frac{\delta F_k}{\delta t}(T|X,Z)^{1_{[\varepsilon=k]}}\right\} \times S(T|X,Z)^{1_{[\varepsilon=0]}} \\
&= \prod_{i=1}^{K}\left\{\lambda_k(T|X,Z)^{1_{[\varepsilon=k]}}(1-F_k(T|X,Z))^{1_{[\varepsilon=k]}}\right\} \times S(T|X,Z)^{1_{[\varepsilon=0]}}
\end{aligned}$$

We obtain then the following log-likelihood :

$$\begin{aligned}
\log(P(T,\varepsilon|X,Z)) &= \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\log(\lambda_k(T|X,Z))\right\} + \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\log(1-F_k(T|X,Z))\right\} \\
&\qquad + 1_{[\varepsilon=0]}\log(S(T|X,Z)) \\
&= \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\log(\lambda_k(T|X,Z))\right\} - \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\Lambda_k(T|X,Z)\right\} + 1_{[\varepsilon=0]}\log(S(T|X,Z))
\end{aligned}$$

And then, introducing the subdistribution proportional hazard model log-likelihood becomes :

$$\begin{aligned}
\log(P(T,\varepsilon|X,Z)) &= \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\log(\lambda_{k0})(t)\right\} + \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}X + \beta_{kZ}Z)\right\} \\
&\quad - \sum_{k=1}^{K}\left\{1_{[\varepsilon=k]}\Lambda_{k0}(t)\exp(\beta_{kX}X + \beta_{kZ}Z)\right\} + 1_{[\varepsilon=0]}\log(S(T|X,Z))
\end{aligned}$$

Using the Bayes theorem, we finally obtain the likelihood of X given T, ε and Z:

$$\begin{aligned}
\log(P(X|T,\varepsilon,Z) &= \log(P(X|Z))) + \log(P(T,\varepsilon|X,Z)) + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\log(\lambda_{k0})(t)\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}X + \beta_{kZ}Z)\right\} \\
&\quad - \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\Lambda_{k0}(t)\exp(\beta_{kX}X + \beta_{kZ}Z)\right\} + 1_{[\varepsilon=0]}\log(S(T|X,Z)) + const \\
&= \log(P(X|Z)) + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\log(\lambda_{k0})(t)\right\} + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}X)\right\} \\
&\quad - \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\Lambda_{k0}(t)\exp(\beta_{kX}X + \beta_{kZ}Z)\right\} + 1_{[\varepsilon=0]}\log(S(T|X,Z)) + const
\end{aligned}$$

where the constant may depend on ε, T and Z but not on X.

If we suppose that X is a binary variable independent from Z, let write logit$(p(X=1|Z))=\zeta$

$$\begin{aligned}
\text{logit}(p(X=1|T,\varepsilon)) &= \log(p(X=1|T,\varepsilon,Z)) - \log(p(X=0|T,\varepsilon))\\
&= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX})\right\} - \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}\Lambda_{k0}(t)\exp(\beta_{kZ}Z)(\exp(\beta_{kX})-1)\right\}\\
&\qquad + 1_{[\varepsilon=0]}\left\{\log(S(T|1)) - \log(S(T|0))\right\}\\
&= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\log(S(T|1)) - \log(S(T|0))\right\}\\
&= \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\}\\
&\qquad + 1_{[\varepsilon=0]}\left\{\log(1-\sum_{i=1}^{K}F_k(T|1)) - \log(1-\sum_{i=1}^{K}F_k(T|0))\right\}
\end{aligned}$$

In absence of censoring, previous equation implies that $(X|T,\varepsilon)$ depends on the bazeline cumulative subdistribution hazard $\Lambda_{\varepsilon 0}$ but not on the other cumulative hazard. In presence of censoring and using a first degree Taylor series approximation we obtain :

$$\begin{aligned}
\text{logit}(p(X=1|T,\varepsilon)) &\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}F_k(T|0) - \sum_{i=1}^{K}F_k(T|1)\right\}\\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}F_k(T|0) - \sum_{i=1}^{K}F_k(T|1)\right\}\\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\}\\
&\qquad + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\exp(\Lambda_k(T|0)) - \sum_{i=1}^{K}\exp(\Lambda_k(T|1))\right\}\\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\} - 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\{\Lambda_{k0}(t)(\exp(\beta_{kX})-1)\}\right\}\\
&\approx \zeta + \sum_{i=1}^{K}\left\{1_{[\varepsilon=k]}(\beta_{kX}+\beta'_{kX}\Lambda_{k0}(t))\right\} + 1_{[\varepsilon=0]}\left\{\sum_{i=1}^{K}\{\beta''_{kX}\Lambda_{k0}(t)\}\right\}
\end{aligned}$$

In that particular case, we demonstrated that as for the sub-distribution approach, the bazeline cumulative subdistribution hazard Λ_{k0} of all the causes of failure i should be included in the imputation model for X with as an interaction term on the type of event observed. As for cause specific hazard, this could be extend to binary X and continuous X with approximations similar to those performed for cause-specific approach.

We demonstrated that as for the cause-specific approach, the baseline cumulative subdistribution hazard Λ_{k0} of all the causes of failure i should be included in the imputation model for X but as an interaction term with the type of event observed contrary to the cause-specific approach. It implies that for a given missing observation only the baseline cumulative subdistribution hazard of the event shared with missing observation directly impacts in the imputation process except for censored observation. As for the baseline cumulative cause specific hazard, baseline cumulative subdistribution hazards could be replaced by estimates of the cumulative subdistribution hazard via a non parametric estimator of the cumulative incidence ([**?**]).

7. Discussion

Our aim in this paper was to propose a method to deal with missing covariates in the presence of competing risks. We developed an approach based on multiple imputation inspired by a previous work of I. White and P. Royston for classical survival analysis [**?**]. Our choice towards a multiple imputation approach was driven by the ability of this framework to handle complex missing data patterns ([**?**, **?**]). Moreover, this approach was already considered during the 2000's to impute the missing causes of failure when one wants to estimate the cause–specific hazard function ([**?**, **?**]) or the cumulative incidence function, either directly ([**?**, **?**]) or through the Fine and Gray model ([**?**, **?**]).

As far as we know, only a recent paper focused on missing covariate values in a competing risks setting, evaluating the long–term effect of covariates on the basis of a mixture model while a copula model modeling the bivariate distribution of both the missing covariates and the missing data mechanisms ([**?**]). Nevertheless, in our opinion, this work presents two main limits. First, it uses a fully parametric model for the hazard of interest which implies a strong constraint in the choice of the model. Secondly, only one hazard is modelled, while it is hardly recommended in the competing risks settings to analyze all the risks that compte to each other ([**?**]).

We proposed two types of imputation models according to the two type of proportional hazard models possibly chosen as analysis model. With proportional cause–specific hazard, we showed the interests to consider both event indicators and cumulative hazards of all the causes of failure as covariates in the imputation model to reduce the biases whatever the type of missingness mechanism (MCAR and MAR) while for proportional subdistribution hazard one should consider in the imputation model cumulative hazards in as an interaction terms with the type of event indicators considered for the missing value. Nevertheless, it appears that cumulative hazards of all the causes of failure should be considered and estimated.

Surprisingly, our work has also an impact on the imputation model that we should use to impute missing covariate in standard survival analysis. In classical survival analysis using Cox proportional hazard model, we consider that inference are valid under the assumption of non informative or independent censoring. In details, it notably means that conditionally on the covariates included in the model, individuals withdrawn from risk at time t should be "representative" of individuals still at risk ([**?**]). Then, it implies that time to censoring could vary with covariates and reversely that time to censoring should possibly contain informations on a missing covariate. When I. White and P. Royston developed their first approach, they considered the classical partial likelihood function. Nevertheless, this partial likelihood is a complete likelihood since it ignores the censoring contributions ([**?**, **?**]). If such an approach is, of course, fully valid to infer on the time to event, it is not so obvious when one want to impute missing covariates. The key concept of multiple imputation is to use information of observed data to estimate plausible value for missing data. But in survival analysis censoring time are also observed data and could also bring information notably when censoring is independent conditionally on covariates. Let consider a classical survival analysis setting in which the outcome is defined by a binary event indicator $D \in \{0, 1\}$ where $D = 1$ denotes the failure and $D = 0$ the right censoring, and a time T to failure or censoring. Using a proportional Cox hazard model or consider a competing risk setting in which the censoring is in competition with the failure and a cause–specific proportional hazard will lead to the same inference. Having this regard on the survival data is only a way to consider globally the information available in the observed data. In such case we showed that we should also introduce the cumulative hazard of the censoring time to impute missing covariates. It is not so unexpected if one want to take into account information linked to the censoring time. Moreover, one can have a congeniality argument to consider cumulative hazard of the censoring time in the imputation model of standard survival data. It is usually considered as crucial in Multiple Imputation that the imputation model should be congenial with the model of interest that will subsequently be fitted to the imputed datasets ([**?**]). Congeniality means that the imputation and analysis models are both compatible with some larger model for the data ([**?**]). In that sense, modelling also the time to censoring is only a way to assure the congeniality of the imputation model and the analysis model by considering a larger model for the imputation which involves the standard survival analysis model. Nevertheless, results obtained by I. White and P. Royston were good in there simulation studies ([**?**]). This is linked to the fact that their way to draw censoring times were independent to the covariates. In practice, such complete independent censoring should always be assume in case of administrative censoring as in a randomized clinical trials. In observational study with censored observations not directly linked to the design of the study, the possible dependence between covariates and censoring times could not be ignored. In such setting, one should consider the cumulative hazard of the censoring time in the imputation model for inferences on standard survival data.

Acknowledgements

xxx