The Performance of Multiple Imputation for Missing Covariate Data within the Context of Regression Relative Survival Analysis

Roch Giorgi[1], Aurélien Belot[2,3], Jean Gaudart[1], Guy Launoy[4] and the French network of cancer registries FRANCIM[5]

[1] Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale, EA 3283, Faculté de Médecine, Université de la Méditerranée, Marseille, France.

[2] Hospices Civils de Lyon, Service de Biostatistique, Lyon ; Université de Lyon ; Université Lyon I ; CNRS UMR 5558, Laboratoire Biostatistique Santé, Pierre-Bénite, France.

[3] Institut de Veille Sanitaire, Département des Maladies Chroniques et des Traumatismes, Saint-Maurice, France.

[4] ERI3 INSERM « Cancers & Populations », FRANCIM, Caen, France.

[5] Head office : Réseau FRANCIM, Faculté de médecine, Toulouse, France.

Corresponding Author:

Roch GIORGI
Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale, EA 3283
Faculté de Médecine, Université de la Méditerranée
27 Boulevard Jean Moulin
F-13385 Marseille Cedex, France
Tel number: 33 (0) 491 324 600
Fax number: 33 (0) 491 324 639
Email: roch.giorgi@univmed.fr

Published in: *Statistics in Medicine* 2008; **27**:6310-6331.

SUMMARY

Relative survival assesses the effects of prognostic factors on disease-specific mortality when the cause of death is uncertain or unavailable. It provides an estimate of patients' survival, allowing for the effects of other independent causes of death. Regression based relative survival models are commonly used in population-based studies to model the effects of some prognostic factors and to estimate net survival. Most often, studies focus on routinely collected prognostic factors for which the proportion of missing values is usually low (around 5%). However, in some cases, additional factors are collected with a greater proportion of missingness. In the present article, we systematically assess the performance of multiple imputation in regression analysis of relative survival through a series of simulation experiments. According to the assumptions concerning the missingness mechanism (completely at random, at random, and not at random) and the missingness pattern (monotone, non-monotone), several strategies were considered and compared: all cases analysis, complete cases analysis, missing data indicator analysis, and multiple imputation by chained equations (MICE) analysis. We showed that MICE performs well in estimating the hazard ratios and the baseline hazard function when the missing mechanism is missing at random conditionally on the vital status. In the situations where the missing mechanism was not MAR conditionally on vital status, complete case behaves consistently. As illustration, we used data of the French Cancer Registries on relative survival of patients with colorectal cancer.

Keywords: missing data; multiple imputation; proportional hazards model; relative survival; colon cancer.

1. INTRODUCTION

Net survival is often part of cancer survival studies. This measure represents survival for a specific disease when all the other causes of death have been removed. Thus, net survival estimates the excess mortality from that disease in the studied group. When the individual causes of death are accurately known (for example, in randomized clinical trials), the analysis of net survival may be carried out with non-disease-related deaths treated as censored observations. However, in cohort studies based on cancer registries, the exact causes of death are often unavailable [1] and, when available, it is often difficult to state whether they are disease-related [2]. In such situations, relative survival is an appropriate approach that provides an estimate of the patients' survival corrected for the effects of other independent causes of death, using the natural mortality in the general population [2]. This method makes it possible to establish whether covariates such as age or sex are associated only with the disease-specific mortality, with the natural mortality in the general population, or with both [3,4]. In other words, relative survival provides a measure of the excess mortality in the patients under study, whether or not the excess mortality is directly attributed to that disease, and whether or not the risks of death from other causes in the studied population are different from those of the general population [5]. This explains why relative survival is popular in population-based cancer survival studies [6] and why relative survival has been an active area of research over the recent decades [e.g., Ref. 6-12].

Although quality control is an important part of cancer registration [13,14], covariate values may be missing for some subjects [15-17]. Most often, population-based cancer survival studies focus on routinely collected prognostic factors such as sex, age, year of diagnosis, tumour size, local or distant metastasis, for which the proportion of missing values is usually low (around 5%). However, for some diseases under study, information on other variables is also collected, and these may have a much higher proportion missing.

A common approach to deal with missing data is to perform a complete case analysis i.e. to exclude patients with any missing value on the covariates or to exclude covariates that are missing in many patients. Such approaches can result in substantially smaller sample sizes and lead to inconsistent estimators of regression coefficients. There is already a large body of work concerning the problem of statistical analysis with missing data and several methods have been proposed to impute missing values before analysis [18-20]. Among those methods, multiple imputation (MI) incorporates uncertainty associated with the missing data appropriately into the subsequent inferences. It is frequently used and several reviews have described its key theoretical ideas, its mode of implementation and compared software implementations [21-25]. Besides, other authors have used MI methods or proposed approaches to handle missing data in regression analysis of crude survival [26-29]. Nevertheless, to the best of our knowledge, the impact of missing data and the assessment of MI have never been studied in the framework of the regression analysis of relative survival.

In this article, we introduce and systematically assess the performance of MI in regression analysis of relative survival. Comparisons are also made with complete case analysis and the use of a missing data indicator. A number of methods that have been proposed for MI are implemented in standard statistical software [23,24]. As regression analysis of relative survival can be performed easily [12,30,31] using the free software R [21], we have chosen to assess the MI by chained equations as available in the multiple imputation by chained equations (MICE) library for R [32]. Section 2 describes our motivating example. This features a population-based dataset with some missing values. Section 3 presents the relative survival regression model and describes briefly the missing data nomenclature and the MICE method. Section 4 describes how we use a simulation study to assess MICE in the regression analysis of relative survival. The motivating example is revisited in section 5 and section 6 concludes with a brief discussion about the findings of the study and recommendations for practical use of this method in the framework of the regression analysis of relative survival.

2. MOTIVATING EXAMPLE

The FRANCIM network is an association that joins all French cancer registries. Among the studies initiated by FRANCIM, high-resolution studies focus on several particular types of medical information that is not routinely collected by cancer registries (such as stage at diagnosis, treatments, socioprofessional category, recurrence and so on). A high-resolution study on colorectal cancer was initiated in 2001, which included 1398 incident cancers diagnosed between 1 January 1995 and 31 December 1995 treated by curative surgery, and identified by 9 French cancer registries (Bas-Rhin, Calvados, Côte d'Or, Doubs, Hérault, Isère, Manche, Somme, Tarn). In our analysis, the follow-up of individual patients was restricted to the first five years after diagnosis, at which time the patients were censored if still alive.

The covariates used for our example, were: sex, age at diagnosis, stage at diagnosis (according to Duke's classification), tumor location, adjuvant treatments, socioprofessional category, and marital status. The other covariates collected for the study were not considered for this analysis. Table I shows the distributions of these prognostic factors. Missing values were observed in 70 patients (5.0%) for the stage at diagnosis, in 29 patients (2.1%) for the adjuvant radiotherapy, in 34 patients (2.5%) for the adjuvant chemotherapy, in 566 patients (40.5%) for the socioprofessional category and in 417 patients (29.8%) for the marital status.

It has been shown that some socioprofessional characteristics may have an impact on survival of patient with colorectal cancer [33,34]. Due to the relatively high proportion of missing values among socioprofessional category and marital status, a complete case regression analysis involving these variables uses only 745 subjects, a reduction of 53.3% from the original 1398. Clearly, this could lead to non-negligible loss of efficiency, in addition to being a potential source of bias.

3. RELATIVE SURVIVAL REGRESSION MODEL AND MULTIPLE IMPUTATION METHOD

3.1. Relative survival regression model

According to the excess hazard model formulation [7,8] of a relative survival regression model, the observed hazard for total mortality, λ_o, at time *t* after diagnosis of an individual aged a at diagnosis and given a vector of covariates $\mathbf{z}$, which could contain age, is defined as the sum of two components:

$$\lambda_o\left(t,\mathrm{z},a\right)=\lambda_e\left(t+a,\mathrm{z}_s\right)+\lambda_c\left(t,\mathrm{z}\right)$$

The first component, λ_e, represents the expected hazard function for overall mortality in the general population. This information is usually obtained from relevant mortality statistics using external sources and depends only on z_s (generally sex, and possibly other factors such as the place of residence, race,…).

The second component, λ_c, represents the disease-related mortality hazard function or the excess mortality hazard, which is estimated from the data on hand and is usually expressed in a parametric form. In the very first proposed models [7,8] λ_c was written $\lambda_c\left(t,\mathbf{z}\right)=\lambda_b\left(t\right)\exp\left(\boldsymbol{\beta}\mathbf{z}\right)$ where $\boldsymbol{\beta}$ is a vector containing the log hazard ratios (HR) of the covariates and λ_b, which represents the baseline excess mortality hazard function (which corresponds to the excess hazard at time t for patients with $\mathbf{z} = 0$), is a step function (i.e. the baseline excess mortality hazard is assumed constant within pre-specified intervals). To allow more flexibility of the baseline excess hazard, some authors have proposed the use of regression splines [9,12,35] and fractional polynomials [11]. Here, we adopt the approach proposed by Remontet *et al.* [12], which uses a cubic regression spline to model the baseline excess hazard. In the absence of covariates, $\log\left[\lambda_c\left(t\right)\right]$ may be written using a truncated power basis:

$$\log\left[\lambda_c\left(t,\mathbf{z}\right)\right] = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \sum_{j=1}^{k} \theta_j \left(t - t_j\right)_+^3$$

where the $'+'$ subscript corresponds to $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$. Expressed in this form, $\log\left[\lambda_c\left(t\right)\right]$ is a smooth piecewise cubic polynomial function, i.e. a smooth combination of subsequent polynomials function in which the function and its first two derivatives are continuous at the knots [36].

3.2. Multiple imputation method

3.2.1. Missing data nomenclature. Consider a sample of n subjects for which we have a vector of response $\mathbf{y}$ and a vector of covariates $\mathbf{z}$. For any given subject, $\mathbf{z}$ may be divided into $\mathbf{z}_{\text{obs}}$, a component to denote the observed covariates and $\mathbf{z}_{\text{mis}}$, a component to denote the missing covariates (to simplify our notation, subject-specific indicator has been omitted). Denote the matrix of missing data indicators by $\mathbf{r}$, with $r_j = 1$ if the jth element of $\mathbf{z}$ is observed, and $r_j = 0$ otherwise, governed by the parameter ϕ.

According to Little and Rubin [20], three missing data mechanisms may be considered: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

- Data are MCAR when the distribution of missingness is independent of $\mathbf{z}_{\text{obs}}$, of $\mathbf{z}_{\text{mis}}$, and on the response $\mathbf{y}$: $\text{P}\left(\mathbf{r}|\mathbf{y},\mathbf{z}\right)=\text{P}\left(\mathbf{r}|\mathbf{y},\mathbf{z}_{\text{obs}},\mathbf{z}_{\text{mis}}\right)=\text{P}\left(\mathbf{r}|\phi\right)$
- Data are MAR when the distribution of missingness is conditionally independent of $\mathbf{z}_{\text{mis}}$ (i.e. when it depends only on observed covariates): $\text{P}\left(\mathbf{r}|\mathbf{y},\mathbf{z}\right)=\text{P}\left(\mathbf{r}|\mathbf{y},\mathbf{z}_{\text{obs}},\phi\right)$

- Data are MNAR when the distribution of missingness $\mathrm{P}(\mathbf{r}|\mathbf{y},\mathbf{z})$ cannot be simplified (i.e. when it depends conditionally on the unobserved covariates). In this case, the process is also termed non-ignorable in the context of a likelihood analysis.

3.2.2. Multiple imputation by chained equations method. Briefly, MI includes three stages: (i) using an assumption of MAR, Bayesian imputations of the missing values are made for M "completed" datasets; (ii) each of these "completed" datasets is analysed; (iii) the estimates and their standard errors are combined using Rubin's rules [37] to produce a single set of estimates with their standard errors.

The MICE imputation algorithm [38] is based on a set of univariate imputation models: the conditional model for each covariate given all the other covariates and the response. The algorithm has two steps: an estimation step and an imputation step. In the estimation step, an imputation model is separately specified for each covariate, involving the other covariates as predictors. In the imputation step, an imputation is generated for the missing variable, and this imputed value is used for the imputation of the next covariate. Using a Gibbs type sampling procedure [39], these steps are repeated until the process is assumed to approach convergence. This is repeated for each of the M datasets.

The analyses were carried out using the R software package (version 2.5.0, April 2007) [40] with the MICE library (version 1.15, March 2007) [32] for multiple imputation.

4. SIMULATION STUDY

In the following simulation, we examined the performance of the MICE method in our relative survival setting. Our strategy consisted of: (i) the generation of all-cases survival data; (ii) the generation of missing values on covariates under different missing values mechanisms and different missingness patterns; (iii) the use of different methods of analysis.

4.1. Survival data generation

Individual survival times were generated from the inverse function of the sum of two additive hazards: the expected hazard function in the general population and the disease-related mortality hazard function. The former was based on the French general population, based on age, and sex. The latter was expressed in a parametric form where the baseline hazard had a generalized Weibull distribution [41], which, as will see later, offers an attractive choice to represent real data-based baseline hazards, and where covariates were sex, age and a vector of other covariates $\mathbf{z}$. The values of the prognostic factors were assumed to be mutually independent. Gender was generated from a binary distribution with $\mathrm{P}(\text{male}) = \mathrm{P}(\text{female}) = 0.5$. Age at diagnosis was categorized into 3 groups: < 65 years, 65-74 years, and ≥ 75 years with probabilities equal to 0.25, 0.35, and 0.40, respectively. Two situations were considered for $\mathbf{z}$: (a) a single binary covariate with $\mathrm{P}(\text{z=1}) = \mathrm{P}(\text{z=0}) = 0.5$; and (b) with 3 binary components, z_1, z_2, and z_3 with $\mathrm{P}(\text{z.=1})$ equal to 0.65, 0.50, and 0.35, respectively.

Individual censoring times were generated in the same way with the hazard selected so as to obtain approximately 15%, 30% or 50% overall censoring levels. Then, an individual's observed time was determined as $T_i = \min(S_i, C_i)$, where S_i and C_i denote the individual's survival and censoring time, respectively.

We used the method proposed by Burton *et al*. [42] to determine the number of simulations required to ignore sampling variation. With an estimated variance from fitting a single covariate in our relative survival regression model equals to 0.01, 584 simulations would be required to produce an estimate to within 2 percentage points of the regression coefficient of $\ln(1.5) = 0.405$ (which corresponds to the minimum of the true values of the log hazard ratios of the covariates) with a 5 per cent significant level. Therefore, each simulation run consisted of 600 independent samples of sizes 300 and 1000. These samples constituted our originals simulated datasets; i.e. the full dataset containing all the cases. The next procedure consisted in generating missing values on the covariates.

4.2. Missing value generation

Several settings were considered to generate missing values on covariates. First, a simple setting with a single binary covariate, z. The missing values were generated according to the three conventional missing value mechanisms: (i) MCAR, for which the probability that z is missing was not linked to any other characteristics; (ii) MAR, for which the probability that z is missing was linked only to sex or only to age or only to vital status or to vital status and sex, or to vital status and age ($\mathrm{P}\left(z_{mis}\right)=0.3$ in males, in patients aged ≥ 65 years at diagnosis, in censored individuals, in censored men, and in censored individuals aged 65 years or older at diagnosis and $\mathrm{P}\left(z_{mis}\right)=0.1$ otherwise); and finally (iii) MNAR, for which the probability that z is missing depended on z itself ($\mathrm{P}\left(z_{mis}\right)=0.8$ when $z=1$ and $\mathrm{P}\left(z_{mis}\right)=0.2$ otherwise). In all these situations, in addition to the individuals' probabilities of missing in the MAR and MNAR situations, the overall proportion of missing values for z was fixed to 10%, 20%, and 50%.

Secondly, we considered more complex situations with three independent binary covariates, z_1, z_2, z_3, and different missingness patterns [18]: (a) a monotone missingness pattern where the probabilities of missing values were fixed to 10% for z_1, 20% for z_2, and 50% for z_3, resulting in an overall proportion of missing values equal to 50%; and (b) a non-monotone missingness pattern with different probabilities of missing values for all possible combinations of the three variables, resulting in an overall proportion of missing values equal to 50%. For both monotone and non-monotone pattern, we considered the procedure described before concerning the missing mechanism (MCAR, MAR and NMAR).

4.3. Design of the evaluation

Several analyses, performed using the above-described relative survival regression model (with a cubic regression spline with one interior knot at 1 year to model the baseline excess hazard), were considered to assess the impact of missing data on covariate in regression analysis of relative survival: (i) an analysis based on all cases (no missing data) and that we used for comparative purposes; (ii) an analysis based on complete cases, a subset of the initial data set without missing data; (iii) an analysis based on all the data with a missing data indicator variable; and (iv) an analysis based on multiple imputation, using MICE (with M = 5 MIs).

The criteria used to assess the methods were: (1) the bias of the estimates (i.e. $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}}$ represents the mean of the estimates of the true value $\boldsymbol{\beta}$); (2) the relative bias of the estimates (i.e. $\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)/\boldsymbol{\beta}$); (3) the empirical coverage rates (ECR) (i.e. the proportion of samples in which the 95% confidence interval includes $\boldsymbol{\beta}$); and (4) the mean of the standard error estimates (MSE). We also investigated the estimate of the baseline excess mortality hazard function, assessed by plotting the mean of 600 estimates of the baseline excess mortality hazard function.

4.4. Results

4.4.1. Missing data on a single binary covariate. Table II summarizes the results obtained with samples of size 1000 and 30% overall censoring level. Without missing values (all cases situation), the model performed relatively well, with a small bias (-0.006), a type I error rate equals to 5.5% and a MSE equals to 0.092. With the complete case analysis, the results were also good in the MCAR, the MAR conditionally on sex (MAR_1) or conditionally on age (MAR_2) and the MNAR situations. However, when the probability of missing was linked to the vital status (MAR_3, MAR_4, and MAR_5), both the bias and type I error increased with increasing proportions of missing data. The results were similar when the analysis used a missing value indicator: the performances were better in the MCAR, MAR_1, MAR_2 and NMAR situations than in the MAR_3, MAR_4 and MAR_5 situations. In those situations in which missingness depends on the vital status (MAR_3, MAR_4 and MAR_5), the MICE method improved the behaviour of the analyses whatever the proportion of missing values (10%, 30%, or 50%): a bias of less than -0.060, an ECR ranging from 91.7% to 95.2% and a MSE less than 0.140.

When the proportion of missing values was 10%, the bias in the estimate of the log hazard ratio was minor in both the complete case and the missing data indicator analyses (Table II), but as previously mentioned, we are also interested in the estimate of the baseline excess mortality hazard function. Figure 1 shows the true baseline excess mortality hazard function (generated by a generalized Weibull distribution, and representing a clinically plausible shape of hazard with increasing risk during the first year after diagnosis, which then decreases as it is often the case in breast cancer) and the mean of the estimates obtained from the different methods used for the analysis with missing values generated under the MAR assumption conditionally on vital status (MAR_3). When the proportion of missing values was about 10% (Figure 1(a)), the estimates obtained with the MICE analysis were similar to those obtained with all cases analysis (the reference) and the estimates obtained with both complete case and missing data indicator analyses were only slightly biased. When the proportion of missing values increased from 10% to 50%, the estimates obtained with the MICE analysis remained unbiased, whereas the bias increased strongly with both complete case and missing data indicator analyses (Figure 1(b)). The same pattern of bias in the baseline hazard function was observed with the other missing data mechanisms associated with vital status (MAR$_4$, MAR$_5$). Concerning the estimates of the baseline hazard function we observe that: (i) in the MCAR, MAR$_1$, and MAR$_2$ situations and whatever the method used for the analysis, the estimates were similar to those obtained with the all case analysis; (ii) in the NMAR situation, the estimates obtained with both the complete case and the missing data indicator analyses were close to those obtained in the all cases analysis, and the estimates obtained with the MICE analysis, wich assumes MAR mechanism, were biased (data not shown).

4.4.2. Missing data on independent binary covariates with different missingness patterns. Table III and Table IV show the bias of the estimates of the log hazard ratios obtained with samples of size 1000 and with a 30% overall censoring level when the missingness pattern was monotone and non-monotone, respectively. Whatever the missingness pattern, complete case analysis overestimated the log HRs in the MCAR, MAR$_1$, MAR$_2$ and NMAR situations and underestimated them in the MAR situations linked to the vital status (MAR$_3$, MAR$_4$ and MAR$_5$). Both the missing data indicator and MICE analyses provided underestimates of the log HRs whatever the missingness mechanism.

When the missingness pattern was monotone, the results showed that: (i) both the bias and the relative bias increased along with the proportion of missing values (i.e. in function of the binary covariates of interest z_i, i=1,…,3); (ii) the bias was larger with the missing data indicator analysis in the MAR$_3$ situation; (iii) MICE analysis performed better than the missing data indicator analysis in the MAR$_3$ situation but comparably in the MAR$_4$ and the MAR$_5$ situations; (iv) the complete case analysis performed well (Table III). When the missingness pattern was non-monotone, the performance of the complete case analysis and the missing data indicator analysis was better than that of MICE (Table IV).

Figure 2 shows that, especially with the monotone missingness pattern, the complete case analysis and the missing data indicator analysis overestimated the baseline excess mortality hazard. This bias was less important with MICE whatever the missingness pattern (Figures 2(a) and 2(b)).

When the simulations summarized in Tables II, III and IV, and in Figures 1 and 2 were repeated with a sample size reduced from 1000 to 300, a generally similar pattern of results was observed (data not shown). All these simulations were also performed with 10% and 50% overall censoring levels and, in a general way, we observed that an increase in the censoring level (from 10% to 50%): (i) had no affect on the point estimates of the log HRs of the covariates and on the estimate of the baseline excess mortality hazard; (ii) resulted in an increase of the MSE (data not shown). All the results presented above using MICE were obtained after M = 5 MIs. To assess the impact of the number of MIs, we also performed analyses with M = 15 and M = 30. That increase in the number of MIs did not affect the previous results (data not shown).

5. THE EXAMPLE REVISITED

Table I shows the proportion of missing values for each covariate of the dataset. The overall proportion of missing values in that dataset was 53.3%. Further investigations have shown a non-monotone missingness pattern and, as shown by the simulation studies, the performance of MICE in estimating the log HRs of the covariates seemed to be less good within this context.

All the covariates described in Section 2 were included in the relative survival regression models. Several approaches were used to model the adjusted effect of age (categorized, linear, quadratic or using cubic regression spline). We have finally chosen the approach with age categorized in three groups; first, for statistical reasons (Akaike Information Criterion, AIC: 3046.9), then because this categorization is quite common [9,35]. Table V shows the results obtained with the complete case analysis, the missing data indicator analysis and the MICE analysis (with M = 5 MIs).

Whatever the analysis method, the adjusted effects of stage at diagnosis were significant while the adjusted effects of gender, tumor location, adjuvant radiotherapy, and marital status were not significant. The adjusted effect of age was significant with the complete case analysis and the MICE analysis but only close to significance with the missing data indicator analysis (p = 0.07). Adjuvant chemotherapy had a significant pejorative effect with the missing data indicator analysis and MICE analysis. This was probably due to the fact that, in this unselected population, chemotherapy was administered to the patients with the worst prognoses. Only the effect of the socioprofessional category “Other persons with occupational activity” was found significant with MICE analysis; that of the association of stage at diagnosis and colorectal-specific mortality was even stronger.

All the estimates obtained with the complete case analysis were different from those obtained with the missing data indicator analysis or MICE analysis. There was a trend towards overestimation of the effects of age, of the socioprofessional category and of the marital status, and another trend towards underestimation of the effects of stage at diagnosis, of adjuvant radiotherapy and of adjuvant chemotherapy. Because of the loss of information, all the 95% confidence intervals obtained with the complete case analysis were larger than those obtained with the missing data indicator analysis or MICE analysis.

The use of the missing data indicator and MICE reduced the standard errors. Whatever the proportion of missing values, the estimates based on the missing data indicator analysis were only slightly different from those based on the MICE analysis (absolute differences less or equal to 0.14). While the socioprofessional categories “Clerical and manual workers” and “Other persons with occupational activity” were found significant in the missing data indicator analysis, only the latter remained significant in the MICE analysis.

Figure 3 shows the baseline excess mortality hazard modelled using a cubic regression spline with one interior knot at 1 year after adjustment for all the covariates considered in this example. Within this context of non-monotone missingness pattern, as shown by the simulation studies, the fits of the baseline excess mortality hazard were similar with both the missing data indicator and the MICE analyses, with a high excess mortality hazard that decreased along the first six months after diagnosis. The complete case analysis captured less well that high risk during that period. The 95% confidence intervals of the baseline excess mortality hazard calculated with the three methods overlapped (data not shown for clarity of Figure 3).

6. DISCUSSION

In this article, we assessed MI in regression analysis of relative survival. Simulations have shown that MICE performed well in estimating the HR and the baseline hazard function when the missing mechanism was MAR conditionally on the vital status with missing values on a single binary covariate or on several binary covariates having a monotone missingness pattern. When the missing mechanism was not MAR conditionally on vital status, complete case analysis was fine. In our motivating example, the proportion of missing values ranged from 2.1% to 40.5%, resulting in a 53.3% overall proportion of missing values (the sample size would be reduced from 1398 to 745).

To our knowledge, the effect of missing data on covariates in relative survival analysis has never been investigated. Yu and Tiwari [5] used MI in relative survival to deal with the missing cause of failure for some individuals but not to tackle the problem of missing data on covariates. They extended the MI method to relative survival data to estimate the net survival function and obtain estimates for covariates with no missing values.

Our simulations have shown that the estimates obtained with the complete case analysis are valid in the MAR situation. However, the introduction of a dependence between covariate missingness and vital status resulted in biased estimates. Those results are consistent with those of Rathouz [43] who, furthermore, proposed two new missingness mechanisms to take into account the dependence between covariate missingness and failure or censoring time: (i) censoring-ignorable missingness at random in which the missingness mechanism depends on the failure time but not on the censoring process; (ii) the failure-ignorable missingness at random in which the missingness mechanism does not depend on the failure time but on the censoring process [43]. Future simulations and empirical studies will improve the knowledge of those new types of missingness mechanisms within the framework of relative survival.

In the regression model used in this article and according to the proportional hazards (PH) assumption, the covariate effects on the disease-related HR was assumed to be constant over time. Several relative survival regression models have been proposed to relax this assumption [4,9,11,12,35]. With the approach proposed by Remontet *et al.* [12], it is possible to model non-proportional hazards using different non-parametric functions and to determine the best model using AIC. We did not use that approach in our example because the simulations did not focus on the PH assumption; we assumed rather constant effects of all covariates. Doing this, the results of the example and those of the simulation can be interpreted the same way.

Several approaches have been proposed for modelling the baseline excess mortality hazard function: step functions [7,8], spline functions [9,12,35], and fractional polynomials [11]. Simulation studies have shown the ability of such functions to provide smooth fits of the baseline excess mortality hazard. However, one may suggest that it may be more interesting to detect changes, increases or decreases, in the trends of the baseline excess mortality hazard instead of giving a smooth fit. Joinpoint regression is one approach to check whether the linear trend of the baseline excess mortality hazard changes at some joinpoints. In that approach, the number and the location of the joinpoints can be determined using permutation tests [44]. A similar approach consists in modelling the baseline excess mortality hazard using a piecewise linear regression with the number of pieces and the location of the joinpoints based on prior information (for example when structural changes are expected at the onset of a new treatment or a new exposure). In our simulation studies, we investigated a piecewise linear regression for the baseline excess mortality hazard using two and three pieces with joinpoints fixed at 1 and 3 years, respectively (based on the shape of the true baseline excess mortality hazard function) with a sample size equal to 1000 and an overall censoring level ranging from 10% to 30%. Within that context, the performance of the methods we used to deal with missing data on covariates were only slightly modified (data not shown). The use of such models depends on the objective of the study: detection of changes in the trend or improvement of the fit of the baseline excess hazard function. However, the goal of this work was not to a deep investigation of such a methodology; thus, further investigations are necessary to evaluate the joinpoint regression in the framework of relative survival.

In our example, we did not use a strategy for variable selection but focused on a full model approach with selected covariates [45]. As in our simulation studies, the baseline excess mortality hazard was modelled using a cubic regression spline, which is probably not the best choice for this dataset (see Figure 3). However, our aim was not to obtain the best fit but to compare the results obtained with different methods that deal with missing data, and, using the same dataset analysis strategy, to obtain results that could be compared to those obtained by simulation. Comparing our results to those obtained by Bolard *et al*. [35] after analysis of relative survival from colon cancer cases extracted from the French

"Bourguignon" registry, we found similar estimates for age and for sex. However, the performance of the method that involves variable selection needs further investigation.

One limitation of our simulation studies is that we focused on binary covariates and we did not investigate missing data on continuous covariates, with possible nonlinear and/or non proportional effect. MICE library supplies several imputation models for continuous covariates: Bayesian linear regression, predictive mean matching and unconditional mean imputation [32]. Further work is needed to assess the performance of MICE in relative survival regression model when the missing values concern continuous covariates.

In this article, MI was assessed only through the use of MI by chained equations method [38] using MICE R library [32]. Several studies performed within other contexts than relative survival regression have assessed different software packages that implement MI, using different methods [22-24]. They found very close results whatever the software package and the analysis methods. In case of missing covariate data in a risk model with a binary outcome, MICE was reported to be the best MI method [46]. MICE and these other methods and softwares might be proposed to analyse missing covariate data within the context of regression analysis of relative survival.

To conclude, in regression analysis of relative survival, missing data on covariates should be modelled. In everyday practice, MICE method offers an attractive choice.

ACKNOWLEDGEMENT

The cancer registries of the French network FRANCIM that participated to the high-resolution study on colorectal cancer are: Bas-Rhin General Cancer Registry (M. Velten), Calvados Digestive Cancer Registry (G. Launoy), Côte d'Or Digestive Cancer Registry (J. Faivre), Doubs General Cancer Registry (A. Danzon), Hérault General Cancer Registry (B. Trétarre), Isère General Cancer Registry (M. Colonna), Manche General Cancer Registry (S. Bara), Somme General Cancer Registry (N. Bourdon-Raverdy), Tarn General Cancer Registry (P. Grosclaude, M. Sauvage).

The authors wish to thank Jean Iwaz for editing the final manuscript.

REFERENCES

1. Percy CL, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. *American Journal of Public Health* **1981**; 71: 242-250.
2. Berkson J, Gage RP. Calculation of survival rates for cancer. *Proceedings of the Staff Meetings of the Mayo Clinic* **1950**; 25: 270-286.
3. Monnet E, Boutron MC, Arveux P, Milan C, Faivre J. Different multiple regressions models for estimating survival: use in a population-based series of colorectal cancer.

Journal of Clinical Epidemiology **1992**; 45: 267-273, DOI:10.1016/0895-4356(92)90086-3.

4. Bolard P, Quantin, C, Esteve J, Faivre J, Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival. Application to colon cancer. *Journal of Clinical Epidemiology* **2001**; 54: 986-96.
5. Yu B, Tiwari RC. Multiple imputation methods for modeling relative survival data. *Statistics in Medicine* **2006**; 25(17): 2946-55, DOI: 10.1002/sim.2289.
6. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* **2004**; 23: 51-64.
7. Hakulinen T, Tenkanen L. Regression analysis of survival rates. *Applied in Statistics* **1987**; 36: 309-317.
8. Esteve J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* **1990**; 9: 529-538.
9. Giorgi R, Abrahamowicz M, Quantin C, Bolard P, Esteve J, Gouvernet J, Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* **2003**; 22: 3767-3784, DOI:10.1002/sim.1484.
10. Stare J, Pohar M, Henderson R. Goodness of fit of relative survival models. *Statistics in Medicine* **2005**; 24(24): 3911-25.
11. Lambert PC, Smith LK, Jones DR, Botha JL. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* **2005**; 24(24): 3871-85.
12. Remontet L, Bossard N, Belot A, Esteve J; French network of cancer registries FRANCIM. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* **2007**; 26(10): 2214-28.
13. Parking DM, Chen VW, Ferlay J, Galceran J, Storm HH, Whelan SL. Comparability and Quality control in Cancer Registration. Technical Report No. 19. Lyon: International Agency for Research on Cancer, **1994**.
14. Tyczynski JE, Démaret E, Parkin DM. Standards and Guidelines for Cancer Registration in Europe. Technical Report No. 40. Lyon: International Agency for Research on Cancer, **2003**.
15. de Vries E, Bray FI, Eggermont AM, Coebergh JW; European Network of Cancer Registries. Monitoring stage-specific trends in melanoma incidence across Europe reveals the need for more complete information on diagnostic characteristics. *European journal of cancer prevention* **2004**; 13(5): 387-95.
16. Threlfall T, Wittorff J, Boutdara P, Heyworth J, Katris P, Sheiner H, Fritschi L. Collection of population-based cancer staging information in Western Australia – a feasibility study. *Population Health Metrics* **2005**; 3: 9, DOI: 10.1186/1478-7954-3-9.
17. Green J, Watson J, Roche M, Beral V, Patnick J. Stage, grade and morphology of tumours of the colon and rectum recorded in the Oxford Cancer Registry, 1995-2003. *British Journal of Cancer* **2007**; 96(1): 140-2, DOI:10.1038/sj.bjc.6603499.
18. Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall: London, **1997**.
19. Allison PD. *Missing Data*. Sage Publications Inc.: Beverley Hills, CA, **2001**.
20. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, second edition. Wiley: New York, **2002**.
21. Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research*

1999; 8: 3-15.

22. Horton NJ, Lipsitz SR. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician* **2001**; 55(3): 244-54.
23. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* **2007**; 61(1): 79-90.
24. Harel O, Zhou XH. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine* **2007**; 26(16): 3057-77.
25. Kenward MG, Carpenter J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* **2007**; 16(3): 199-218.
26. Paik MC. Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis* **1997**; 3(3): 289-98.
27. Clark TG, Stewart ME, Altman DG, Gabra H, Smyth JF. A prognostic model for ovarian cancer. *British Journal of Cancer* **2001**; 85(7): 944-52.
28. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. *American Journal of Epidemiology* **2003**; 157(1): 74-84.
29. Aitkin I, Scott J. The effect of missing data on covariates in survival analysis. Stata User Group - 1st Australian and New Zealand Users Group meeting **2004**. Available online at http://www.stata.com/meeting/1australia/ [accessed August 29, 2007].
30. Giorgi R, Payan J, Gouvernet J. RSURV: a function to perform relative survival analysis with S-PLUS or R. *Computer Methods and Programs in Biomedicine* **2005**; 78: 175-178, DOI:10.1016/j.cmpb.2005.01.001.
31. Pohar M, Stare J. Relative survival analysis in R. *Computer Methods and Programs in Biomedicine* **2006**; 81(3): 272-8.
32. Van Buuren S. Multiple Imputation Online **2007**. Available online at http://www.multiple-imputation.com [accessed August 29, 2007].
33. Sapp AL, Trentham-Dietz A, Newcomb PA, Hampton JM, Moinpour CM, Remington PL. Social networks and quality of life among female long-term colorectal cancer survivors. *Cancer* **2003**; 98(8): 1749-58.
34. Dejardin O, Remontet L, Bouvier AM, Danzon A, Tretarre B, Delafosse P, Molinie F, Maarouf N, Velten M, Sauleau EA, Bourdon-Raverdy N, Grosclaude P, Boutreux S, De Pouvourville G, Launoy G. Socioeconomic and geographic determinants of survival of patients with digestive cancer in France. *British Journal of Cancer* **2006**; 95(7): 944-9.
35. Bolard P, Quantin C, Abrahamowicz M, Esteve J, Giorgi R, Chadha-Boreham H, Binquet C, Faivre J. Assessing time-by-covariate interactions in relative survival models using restrictive cubic spline functions: application to colon cancer. *Journal of Cancer Epidemiology and Prevention* **2002**; 7:113-122.
36. de Boor. A Practical Guide to Splines. Springer: New York, **1978**.
37. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York, **1987**.
38. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **1999**; 18(6): 681-94.
39. Geman S, Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence* **1984**; 6: 721-741.
40. R Development Core Team, R. A language and Environment for Statistical Computing,

R Foundation for Statistical Computing, Vienna, Austria, **2007**. Available online at http://www.R-project.org [accessed August 29, 2007].

41. Mudholkar GS, Srivastava DK, Kollia GD. A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. *Journal of the American Statistical Association* **1996**; 91(436): 1575-1583, DOI:10.2307/2291583.
42. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* **2006**;25:4279:4292.
43. Rathouz PJ. Identifiability assumptions for missing covariate data in failure time regression models. *Biostatistics* **2007**; 8(2): 345-56, DOI:10.1093/biostatistics/kx1014.
44. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine* **2000**;19:335:51.
45. Harrell FE. *Regression modeling strategies*. Springer: New York, **2001**.
46. Ambler G, Omar RZ, Royston P. A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research* **2007**; 16(3): 277-98.

Table I: Distributions of the colorectal cancer prognostic factors and the corresponding all-causes mortality in a population-based study of French colorectal cancer.

Prognostic Factors	Number (%) [a]	Deaths at 5 years (%) [b]
Age		
≤ 64	392 (28.0%)	90 (23.0%)
65 – 74	488 (34.9%)	156 (32.0%)
≥75	518 (37.1%)	282 (54.4%)
Gender		
Man	755 (54.0%)	297 (39.3%)
Woman	643 (46.0%)	231 (35.9%)
Tumor location		
Colon	854 (61.1%)	325 (38.1%)
Rectosigmoid or rectum	544 (38.9%)	203 (37.3%)
Tumor Stage at diagnosis		
A	395 (28.2%)	68 (17.2%)
B	495 (35.4%)	198 (40.0%)
C	395 (28.2%)	214 (54.2%)
D	43 (3.1%)	32 (74.4%)
Missing	70 (5.0%)	16 (22.9%)
Adjuvant radiotherapy		
No	225 (16.1%)	92 (40.9%)
Yes	1144 (81.8%)	427 (37.3%)
Missing	29 (2.1%)	9 (31.0%)
Adjuvant chemotherapy		
No	392 (28.0%)	142 (36.2%)
Yes	972 (69.5%)	374 (38.5%)
Missing	34 (2.5%)	12 (35.3%)
Socioprofessional category		
Farmers	134 (9.3%)	46 (35.4%)
Clerical and manual workers	214 (15.3%)	82 (38.3%)
Other with occupational activity	360 (25.7%)	129 (35.8%)
Other without occupational activity	128 (9.2%)	42 (32.8%)
Missing	566 (40.5%)	229 (40.5%)
Marital status		
Married or living with a partner	359 (25.7%)	153 (42.6%)
Single or widowed	622 (44.5%)	221 (35.5%)
Missing	417 (29.8%)	154 (36.9%)
Overall	1398 (100%)	528 (37.8%)

[a] Percentage of all 1398 patients.

[b] Percentage of patients, in a given category, who died within the first 5 years after diagnosis.

Table II: Results of simulation studies for the all cases analysis, and, when missing data concerned a single binary covariate z, for the complete cases analysis, the analysis using a missing data indicator, and the analysis using MICE (simulation size = 600, sample size = 1000, and overall censoring level = 30%). Results are shown when the true value of $\beta_z = \ln(2.0)$ and are adjusted for sex and age.

Missing mechanism [a]	All Cases	Complete cases			Missing data indicator			Multiple Imputation		
		Proportion of missing values for z			Proportion of missing values on z			Proportion of missing values on z		
		10%	30%	50%	10%	30%	50%	10%	30%	50%
MCAR										
Bias	-0.006	-0.007	-0.007	0.001	-0.009	-0.015	-0.006	-0.019	-0.044	-0.055
ECR	94.5	94.3	93.5	93.3	94.5	94.3	94.2	93.7	91.7	92.3
MSE	0.092	0.098	0.111	0.132	0.097	0.110	0.130	0.099	0.113	0.133
MAR_1										
Bias	-0.006	-0.008	-0.005	-0.007	-0.010	-0.013	-0.022	-0.018	-0.032	-0.049
ECR	94.5	93.2	95.2	94.8	93.5	94.8	94.8	93.2	93.2	92.3
MSE	0.092	0.097	0.110	0.130	0.097	0.109	0.128	0.098	0.112	0.131
MAR_2										
Bias	-0.006	-0.006	-0.003	0.010	-0.008	-0.010	-0.003	-0.025	-0.059	-0.094
ECR	94.5	94.5	95.2	95.0	94.5	95.2	95.3	93.5	92.1	87.7
MSE	0.092	0.098	0.111	0.133	0.097	0.111	0.131	0.098	0.112	0.128
MAR_3										
Bias	-0.006	-0.026	-0.073	-0.121	-0.028	-0.081	-0.135	-0.013	-0.025	-0.022
ECR	94.5	94.0	87.5	81.7	93.8	86.2	78.7	95.2	92.8	92.3
MSE	0.092	0.095	0.103	0.115	0.095	0.102	0.114	0.098	0.114	0.140
MAR_4										
Bias	-0.006	-0.019	-0.044	-0.070	-0.020	-0.048	-0.072	-0.013	-0.027	-0.036
ECR	94.5	93.5	92.5	90.5	93.3	91.5	90.2	94.5	92.3	91.7
MSE	0.092	0.096	0.106	0.122	0.096	0.105	0.121	0.098	0.113	0.136
MAR_5										
Bias	-0.006	-0.019	-0.054	-0.092	-0.021	-0.062	-0.104	-0.013	-0.036	-0.060
ECR	94.5	93.0	92.2	87.8	93.0	91.0	86.7	93.8	93.8	92.7
MSE	0.092	0.096	0.105	0.120	0.096	0.104	0.119	0.099	0.114	0.137
MNAR										
Bias	-0.006	-0.006	-0.009	0.002	-0.007	-0.016	-0.011	-0.025	-0.069	-0.104
ECR	94.5	94.5	95.8	94.8	94.0	95.6	94.6	93.7	88.2	87.5
MSE	0.092	0.097	0.112	0.142	0.097	0.111	0.140	0.097	0.109	0.133

[a] MAR_1 = MAR conditionally on sex - MAR_2 = MAR conditionally on age - MAR_3 = MAR conditionally on vital status - MAR_4 = MAR conditionally on vital status and sex - MAR_5 = MAR conditionally on vital status and age.

Abbreviations: ECR: empirical coverage rates - MSE, mean of the standard error estimates.

Table III: Monotone missingness pattern: bias of the estimates of the log hazard ratios for the all cases analysis, and, when missing data concerned three binary covariates (overall proportion of missing values = 50%), for the complete cases analysis, the analysis using a missing data indicator and the analysis using MICE (simulation size = 600, sample size = 1000, and overall censoring level = 30%).

Missing	All Cases			Complete cases			Missing data indicator			Multiple Imputation		
mechanism[a]	Variables[b]			Variables[b]			Variables[b]			Variables[b]		
	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
MCAR												
Bias	0.005	0.007	0.008	0.009	0.016	0.019	-0.032	-0.074	-0.086	-0.048	-0.112	-0.155
Rel. Bias	0.012	0.008	0.006	0.022	0.018	0.015	-0.079	-0.081	-0.069	-0.117	-0.123	-0.124
ECR	95.2	95.5	94.7	95.2	95.2	95.2	93.8	87.3	90.5	94.2	82.3	75.5
MSE	0.090	0.089	0.092	0.129	0.126	0.132	0.095	0.097	0.124	0.106	0.109	0.136
MAR_1												
Bias	0.005	0.007	0.008	0.013	0.013	0.018	-0.032	-0.072	-0.085	-0.045	-0.105	-0.131
Rel. Bias	0.012	0.008	0.006	0.032	0.014	0.014	-0.079	-0.078	-0.068	-0.112	-0.114	-0.105
ECR	95.2	95.5	94.7	95.0	93.7	94.0	95.2	88.3	90.3	95.7	83.0	80.2
MSE	0.090	0.089	0.092	0.128	0.125	0.131	0.095	0.097	0.123	0.105	0.106	0.133
MAR_2												
Bias	0.005	0.007	0.008	0.013	0.013	0.019	-0.034	-0.073	-0.087	-0.053	-0.129	-0.217
Rel. Bias	0.012	0.008	0.006	0.032	0.015	0.015	-0.084	-0.080	-0.069	-0.131	-0.141	-0.173
ECR	95.2	95.5	94.7	95.2	94.5	94.7	93.7	87.8	90.5	94.2	77.7	61.2
MSE	0.090	0.089	0.092	0.130	0.127	0.132	0.095	0.098	0.125	0.106	0.107	0.134
MAR_3												
Bias	0.005	0.007	0.008	-0.031	-0.079	-0.099	-0.037	-0.097	-0.156	-0.033	-0.080	-0.102
Rel. Bias	0.012	0.008	0.006	-0.077	-0.086	-0.079	-0.091	-0.105	-0.125	-0.083	-0.087	-0.081
ECR	95.2	95.5	94.7	94.3	88.3	85.8	93.7	82.5	75.3	95.3	89.2	87.8
MSE	0.090	0.089	0.092	0.114	0.112	0.117	0.093	0.094	0.112	0.103	0.106	0.129
MAR_4												
Bias	0.005	0.007	0.008	-0.013	-0.041	-0.042	-0.034	-0.083	-0.113	-0.041	-0.091	-0.116
Rel. Bias	0.012	0.008	0.006	-0.031	-0.045	-0.033	-0.084	-0.091	-0.090	-0.101	-0.100	-0.093
ECR	95.2	95.5	94.7	94.7	91.8	92.5	92.8	86.7	84.0	94.7	86.5	81.7
MSE	0.090	0.089	0.092	0.121	0.118	0.124	0.094	0.095	0.117	0.104	0.106	0.131
MAR_5												
Bias	0.005	0.007	0.008	-0.021	-0.052	-0.065	-0.035	-0.086	-0.129	-0.038	-0.098	-0.131
Rel. Bias	0.012	0.008	0.006	-0.052	-0.057	-0.052	-0.085	-0.094	-0.103	-0.095	-0.107	-0.105
ECR	95.2	95.5	94.7	94.7	91.2	92.5	93.3	84.7	82.2	95.7	84.8	80.3
MSE	0.090	0.089	0.092	0.119	0.116	0.121	0.094	0.095	0.115	0.105	0.105	0.133
MNAR												
Bias	0.005	0.007	0.008	0.002	0.016	0.019	-0.038	-0.082	-0.094	-0.056	-0.134	-0.215
Rel. Bias	0.012	0.008	0.006	0.006	0.018	0.015	-0.095	-0.089	-0.075	-0.138	-0.146	-0.172
ECR	95.2	95.5	94.7	94.3	95.2	94.0	93.7	86.3	90.0	93.5	71.5	66.3
MSE	0.090	0.089	0.092	0.131	0.130	0.151	0.094	0.097	0.144	0.100	0.103	0.146

Note: The proportion of missing values for z_1, z_2, and z_3 were, fixed to 10%, 20%, and 50%, respectively (overall proportion of missing values = 50%).

[a] MAR_1 = MAR conditionally on sex - MAR_2 = MAR conditionally on age - MAR_3 = MAR conditionally on vital status - MAR_4 = MAR conditionally on vital status and sex - MAR_5 = MAR conditionally on vital status and age.

[b] The true value of $\beta_{z1} = \ln(1.5)$, $\beta_{z2} = \ln(2.5)$, $\beta_{z3} = \ln(3.5)$. The estimates were adjusted for sex and age.

Abbreviations: Rel. Bias: relative bias; ECR: empirical coverage rates; MSE, mean of the standard error estimates.

Table IV: Non-monotone missingness pattern: bias of the estimates of the log hazard ratios for the all cases analysis, and, when missing data concerned three binary covariates (overall proportion of missing values = 50%), for the complete cases analysis, the analysis using a missing data indicator and the analysis using MICE (simulation size = 600, sample size = 1000, and overall censoring level = 30%).

Missing	All Cases			Complete cases			Missing data indicator			Multiple Imputation		
Mechanism[a]	Variables[b]			Variables[b]			Variables[b]			Variables[b]		
	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3	z_1	z_2	z_3
MCAR												
Bias	0.008	0.007	0.008	0.010	0.017	0.014	-0.024	-0.051	-0.071	-0.065	-0.116	-0.120
Rel. Bias	0.019	0.008	0.007	0.026	0.018	0.011	-0.059	-0.056	-0.057	-0.160	-0.127	-0.096
ECR	95.2	95.5	94.7	95.5	94.7	94.5	94.8	93.7	90.0	93.2	81.0	80.0
MSE	0.090	0.089	0.092	0.127	0.125	0.130	0.106	0.103	0.105	0.112	0.110	0.116
MAR_1												
Bias	0.008	0.007	0.008	0.013	0.014	0.024	-0.021	-0.051	-0.065	-0.063	-0.111	-0.106
Rel. Bias	0.019	0.008	0.007	0.032	0.015	0.019	-0.051	-0.056	-0.052	-0.155	-0.121	-0.085
ECR	95.2	95.5	94.7	95.5	94.2	95.0	95.2	92.5	91.8	93.7	83.8	84.3
MSE	0.090	0.089	0.092	0.126	0.124	0.130	0.106	0.102	0.105	0.112	0.110	0.115
MAR_2												
Bias	0.008	0.007	0.008	0.013	0.018	0.018	-0.020	-0.052	-0.073	-0.067	-0.134	-0.139
Rel. Bias	0.019	0.008	0.007	0.032	0.019	0.014	-0.050	-0.057	-0.058	-0.166	-0.147	-0.111
ECR	95.2	95.5	94.7	95.7	94.0	94.0	95.0	92.2	89.7	93.8	78.7	76.0
MSE	0.090	0.089	0.092	0.128	0.125	0.130	0.107	0.103	0.105	0.113	0.110	0.116
MAR_3												
Bias	0.008	0.007	0.008	-0.033	-0.085	-0.099	-0.030	-0.079	-0.096	-0.044	-0.088	-0.089
Rel. Bias	0.019	0.008	0.007	-0.082	-0.093	-0.079	-0.075	-0.086	-0.077	-0.109	-0.096	-0.071
ECR	95.2	95.5	94.7	93.7	86.5	83.8	94.3	87.0	83.8	94.8	87.3	87.2
MSE	0.090	0.089	0.092	0.113	0.111	0.116	0.102	0.099	0.102	0.111	0.108	0.114
MAR_4												
Bias	0.008	0.007	0.008	-0.009	-0.040	-0.049	-0.021	-0.066	-0.082	-0.049	-0.106	-0.107
Rel. Bias	0.019	0.008	0.007	-0.023	-0.043	-0.039	-0.053	-0.072	-0.065	-0.121	-0.116	-0.085
ECR	95.2	95.5	94.7	94.5	93.0	92.8	94.7	88.8	88.5	94.8	82.2	84.0
MSE	0.090	0.089	0.092	0.120	0.117	0.122	0.104	0.101	0.104	0.112	0.108	0.115
MAR_5												
Bias	0.008	0.007	0.008	-0.023	-0.052	-0.062	-0.030	-0.064	-0.081	-0.058	-0.103	-0.100
Rel. Bias	0.019	0.008	0.007	-0.058	-0.057	-0.049	-0.073	-0.070	-0.065	-0.143	-0.112	-0.080
ECR	95.2	95.5	94.7	94.7	91.8	90.7	93.5	88.7	87.5	92.3	85.0	82.5
MSE	0.090	0.089	0.092	0.117	0.115	0.120	0.103	0.100	0.103	0.113	0.108	0.114
MNAR												
Bias	0.008	0.007	0.008	0.002	0.016	0.015	-0.016	-0.048	-0.067	-0.068	-0.124	-0.123
Rel. Bias	0.019	0.008	0.007	0.004	0.017	0.012	-0.038	-0.053	-0.054	-0.167	-0.135	-0.098
ECR	95.2	95.5	94.7	93.3	94.0	94.8	92.7	94.0	90.8	90.8	79.2	82.7
MSE	0.090	0.089	0.092	0.126	0.126	0.134	0.104	0.102	0.107	0.108	0.108	0.116

Note: The different probabilities of missing for all the possible combination of z_1, z_2, and z_3 resulted in an overall proportion of missing values = 50%.

[a] MAR_1 = MAR conditionally on sex, MAR_2 = MAR conditionally on age, MAR_3 = MAR conditionally on vital status, MAR_4 = MAR conditionally on vital status and sex, MAR_5 = MAR conditionally on vital status and age.

[b] The true value of $\beta_{z1} = \ln(1.5)$, $\beta_{z2} = \ln(2.5)$, $\beta_{z3} = \ln(3.5)$. The estimates were adjusted for sex and age.

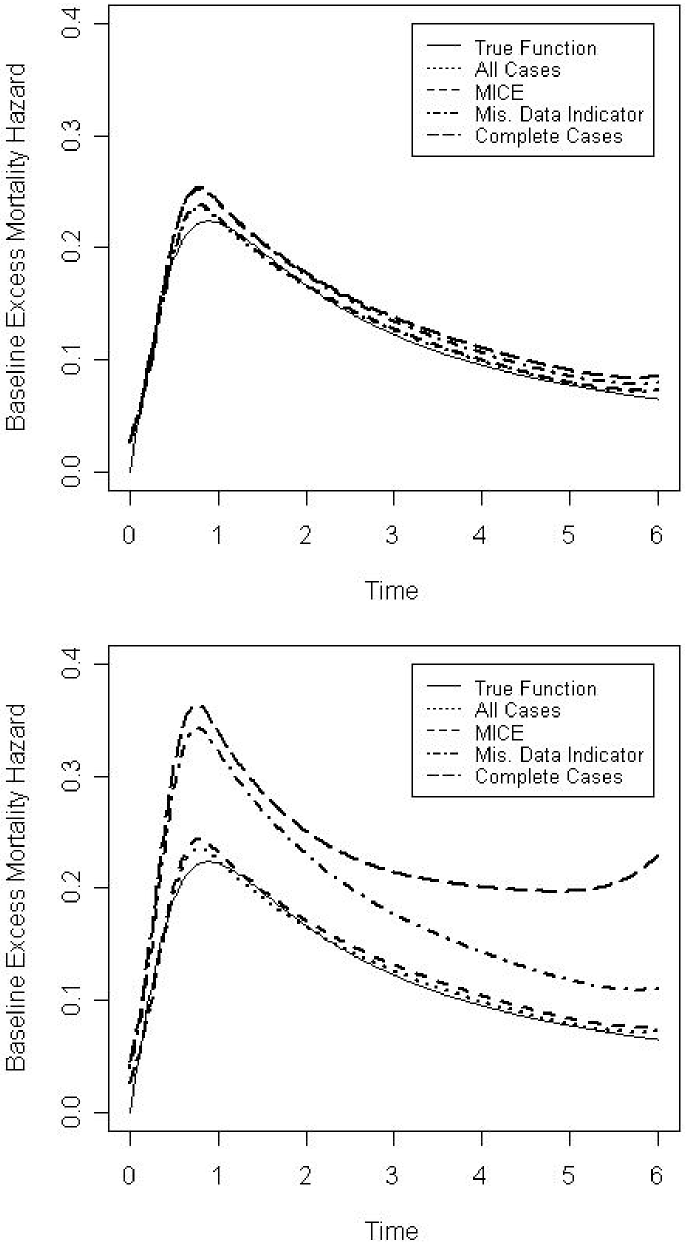
Abbreviations: Rel. Bias: relative bias; ECR: empirical coverage rates; MSE, mean of the standard error estimates.

Table V: Adjusted log hazard ratios (HR), and 95% confidence intervals, obtained by regression analysis of relative survival in a population-based study of French colorectal cancer. Results obtained using the complete case analysis, the missing data indicator analysis and the multiple imputations analysis.

	Complete cases			Missing data indicator			Multiple Imputation		
	N = 745			N=1398			N=1398		
Prognostic Factors[c]	log HR	95% CI	P-values	log HR	95% CI	P-values	log HR	95% CI	P-values
Age									
≤ 64	0.00		0.03	0.00		0.07	0.00		0.04
65 – 74	0.52	0.09 ; 0.95		0.33	0.02 ; 0.64		0.32	0.01 ; 0.63	
≥75	0.69	0.17 ; 1.21		0.42	0.06 ; 0.77		0.48	0.12 ; 0.85	
Gender									
Man	0.00		0.83	0.00		0.39	0.00		0.50
Woman	-0.05	-0.42 ; 0.33		-0.12	-0.37 ; 0.13		-0.10	-0.37 ; 0.17	
Tumor location									
Colon	0.00		0.80	0.00		0.66	0.00		0.66
Rectosigmoid or rectum	-0.07	-0.51 ; 0.38		-0.07	-0.37 ; 0.22		-0.08	-0.37 ; 0.22	
Stage at diagnosis									
A	0.00		<0.001	0.00		<0.001	0.00		<0.001
B	1.69	0.91 ; 2.48		1.99	1.45 ; 2.52		2.07	1.48 ; 2.67	
C-D	2.74	1.96 ; 3.53		2.85	2.31 ; 3.39		2.92	2.33 ; 3.51	
Missing				-0.24	-1.70 ; 1.22				
Adjuvant radiotherapy									
No	0.00		0.77	0.00		0.12	0.00		0.18
Yes	-0.09	-0.61 ; 0.43		-0.30	-0.65 ; 0.05		-0.28	-0.66 ; 0.11	
Missing				-0.61	-2.37 ; 1.14				
Adjuvant chemotherapy									
No	0.00		0.35	0.00		0.004	0.00		0.01
Yes	0.22	-0.20 ; 0.63		0.47	0.18 ; 0.76		0.45	0.14 ; 0.75	
Missing				0.50	-0.90 ; 1.91				
Socioprofessional category									
Farmers	0.00		0.05	0.00		0.06	0.00		0.02
Clerical and manworkers	0.63	0.05 ; 1.22		0.56	0.03 ; 1.09		0.42	-0.17 ; 1.01	
Other with occupational activity	0.79	0.25 ; 1.33		0.67	0.17 ; 1.17		0.69	0.12 ; 1.26	
Other without occupational activity	0.41	-0.28 ; 1.10		0.36	-0.26 ; 0.97		0.42	-0.19 ; 1.04	
Missing				0.63	0.14 ; 1.12				
Marital status									
Married of with partner	0.00		0.63	0.00		0.61	0.00		0.81
Single or widowed	0.11	-0.29 ; 0.51		-0.02	-0.32 ; 0.28		0.01	-0.28 ; 0.29	
Missing				-0.09	-0.41 ; 0.24				

Abbreviations: CI = Confidence Interval – HR: Hazard Ratio.

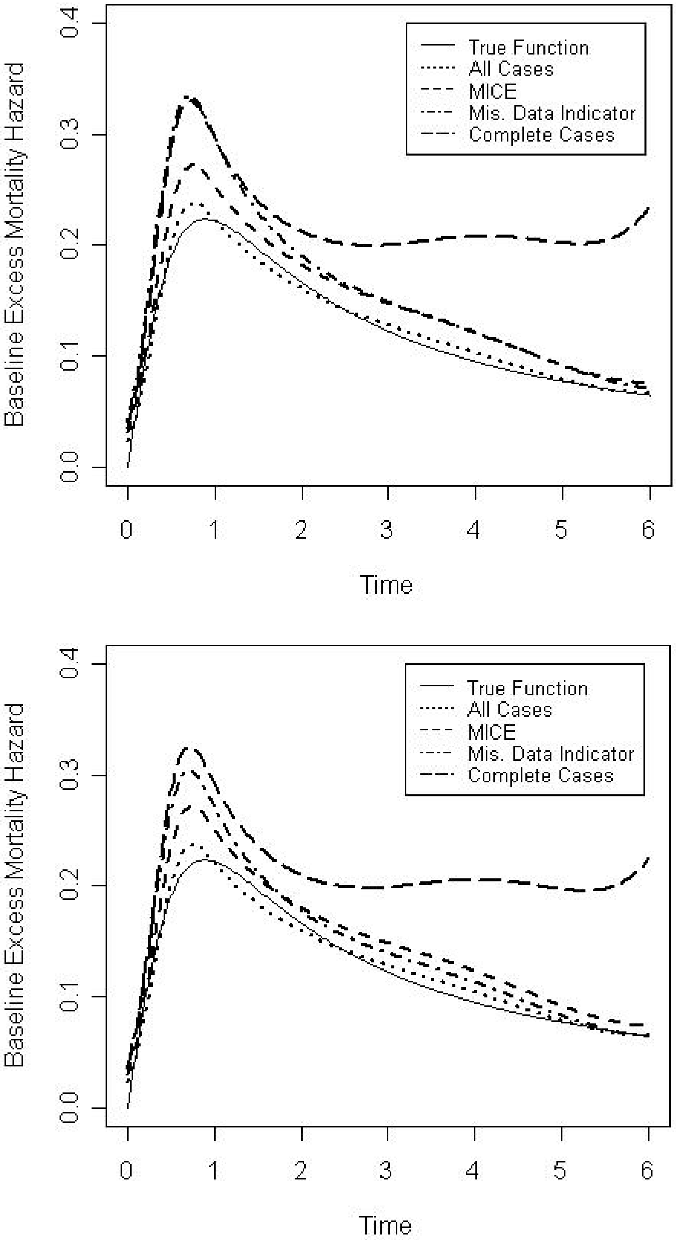
Figure 1. True baseline excess mortality hazard functions against the mean of 600 estimates (sample size = 1000; overall censoring level = 30%) according to the method used to analyze the data when the missing values were generated under the MAR assumption conditionally on the vital status. The overall proportion of missing values was fixed to (a) 10% and (b) 50%.



(a)

(b)

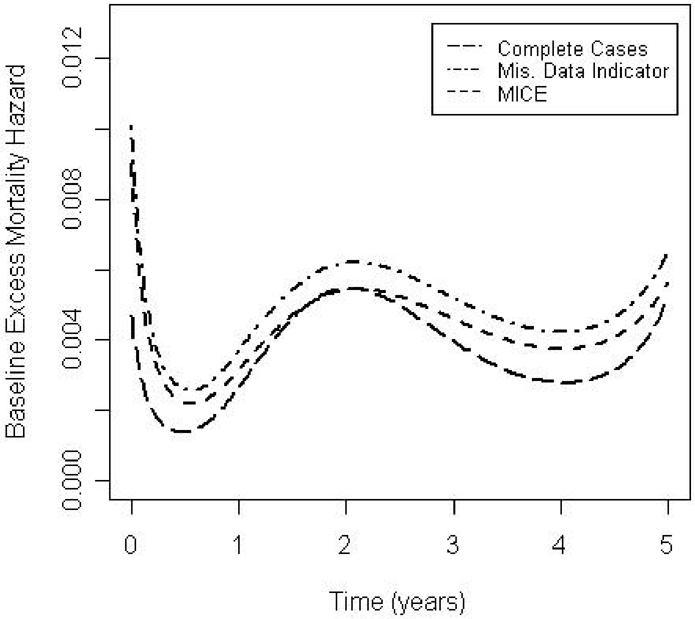
Figure 2. True baseline excess mortality hazard functions against the mean of 600 estimates (sample size = 1000; overall censoring level = 30%) according to the method used to analyze the data when the missing values were generated under the MAR assumption conditionally on the vital status, and when the missingness pattern was (a) monotone and (b) non-monotone.



(a)

(b)

Figure 3. Baseline excess mortality hazard (adjusted for all the covariates analysed; see Table 1), according to the method used to analyze the population-based study of French colorectal cancer.


Complete Cases
Mis. Data Indicator
MICE
Baseline Excess Mortality Hazard
Time (years)