Dealing with missing information on covariates for excess mortality hazard regression models – making the imputation model compatible with the substantive model

Luís Antunes[1,2], Denisa Mendonça[2,3], Maria José Bento[1,3], Edmund Njeru Njagi[4], Aurélien Belot[4], Bernard Rachet[4]

[1] Grupo de Epidemiologia de Cancro, Centro de Investigação do IPO Porto (CI-IPOP), Instituto Português de Oncologia do Porto (IPO Porto), Porto, Portugal

[2] EPI-UNIT - Instituto de Saúde Pública, Universidade do Porto, Porto, Portugal

[3] Instituto de Ciências Biomédicas Abel Salazar, Universidade do Porto, Portugal

[4] Inequalities in Cancer Outcomes Network, London School of Hygiene and Tropical Medicine, United Kingdom

Corresponding author:

Luis Antunes

Rua Dr. António Bernardino de Almeida

4200-072 Porto, Portugal

Email: antunes.lj@gmail.com

Abstract

Missing data is a common issue in epidemiological databases. Among the different ways of dealing with missing data, multiple imputation has become more available in common statistical software packages. However the incompatibility between the imputation and substantive model, which can arise when the associations between variables in the substantive model are not taken into account in the imputation models or when the substantive model is itself nonlinear, can lead to invalid inference. Aiming at analysing population-based cancer survival data, we extended the multiple imputation substantive model compatible fully conditional specification (SMC-FCS) approach, proposed by Bartlett and colleagues in 2015, to accommodate excess hazard regression models. The proposed approach was compared with the standard fully conditional (FCS) multiple imputation procedure and with the complete-case analysis (CCA) using a simulation study. The SMC-FCS approach produced unbiased estimates in both scenarios tested, while the FCS produced biased estimates and poor empirical coverages probabilities. The SMC-FCS algorithm was then used for handling missing data in the evaluation of socioeconomic inequalities in survival from colorectal cancer patients diagnosed in the North Region of Portugal. The analysis using SMC-FCS showed a clearer trend in higher excess hazards for patients coming from more deprived areas.

The proposed algorithm was implemented in R software and is presented as supplementary material.

Keywords: Missing data, multiple imputation, substantive model compatible, excess hazard, socioeconomic inequalities

1. Introduction

Missing data is an almost unavoidable issue in observational studies. Due to multiple possible reasons, incomplete information on the outcomes or on the covariates is likely to occur. Multiple imputation (MI) has in recent years become one of the most common methodologies for handling missing data [1,2]. Its increasing availability in common statistical packages has made the application of MI more attractive to a larger spectrum of users[3]. Unfortunately, this broader application of the methodology has not necessarily been followed by a correct application or reporting of the same. Rezvan and colleagues systematically reviewed manuscripts published during six years in two important medical journals in which multiple imputation was carried out [2]. From the 103 articles identified, only 37% described the imputation model, only two compared the imputed with the observed values and only three performed sensitivity analysis.

Also, the problem of incompatibility between imputation model and the substantive (or analysis) model can lead to invalid inference. This problem can occur when the substantive model includes nonlinear covariate effects, interactions or when the model itself is nonlinear (e.g. hazard models)[4].

When the outcome of interest is survival time and there is missing information on covariates, there is consensus that the outcome should be included in the imputation model. However, different ways of including the survival outcome can be found in the literature: the censoring indicator (δ) and the survival time (T) [5]; δ and *log(T)* [6,7]; δ, *log(T)* and T [8]. In 2009, White and Royston[9] showed that when the substantive model is a Cox hazard model, a suitable model for imputing binary or Normal variables is a logistic or linear regression on the cumulative baseline hazard (Λ_0(t)),the censoring indicator and the other covariates.

In 2015, Bartlett and colleagues developed an algorithm for MI that ensures compatibility between the imputation and substantive model, and named it Substantive Model Compatible Fully Conditional Specification (SMC-FCS) [4]. This method has been implemented in STATA and R but only a limited number of substantive models are available [10]. Later, Keogh and Morris [11] extended this approach to hazard models with time-varying effects of covariates.

In population-based cancer survival analysis, interest typically focuses on estimating cancer-specific quantities within the so-called “relative survival setting” [12,13]. In the relative survival setting, we assume that the overall mortality hazard for a patient *i* may be written as the sum of an expected mortality hazard (as observed in the general population) and an excess mortality hazard. The expected mortality hazard is considered known (as provided by life table) and is considered as an estimate of the other-cause mortality. The interest is in estimating the excess mortality hazard (and the corresponding net survival) as it is assumed that the excess mortality hazard represents the mortality hazard due (directly or indirectly) to the disease under study and is now commonly modelled using flexible parametric regression models [14,15].

Multiple imputation has been used to deal with missing covariate information on excess mortality hazard regression modelling [16–21]. In 2015, Falcaro and colleagues evaluated the use of MI in the context of net survival problems with missing information, more specifically, in the excess hazard modelling using flexible parametric proportional hazards models with missing data on categorical covariates (stage of disease at diagnosis) [22]. The results obtained suggested that a multinomial logistic imputation model for stage should be used and that the Nelson-Aalen cumulative excess hazard estimate and the event indicator should be included in the imputation models, as already suggested by White and Royston in the context of the Cox model. The issue of compatibility between the imputation and substantive models when these are excess hazard models has however still not been properly addressed.

The main aim of this work was to extend the SMC-FCS algorithm developed by Bartlett and colleagues to accommodate excess hazard models. The performance of the extension proposed was compared with the standard fully conditional specification

(FCS) approach and with a complete-case analysis (CCA), using a simulation study. The three methods were then applied to a survival dataset from a cohort of colorectal cancer patients extracted from the North Region of Portugal Cancer Registry (RORENO).

The article is organised as follows. In Section 2, an overview of the methods used in this study is given and the proposed extension of the SMC-FCS algorithm for excess hazard models is presented. A simulation study evaluating the performance of the SMC-FCS algorithm is presented in Section 3. The motivating dataset is introduced and then analysed in Section 4 with the aim of evaluating socioeconomic inequalities in survival from cancer when adjusting for extent of disease at diagnosis. The article concludes with a discussion in Section 5.

2. Methods

This study focus on the occurrence of missing data on covariates in excess hazard models. We start by giving an introduction to the concept of net survival and excess hazard followed by a brief explanation of the type of excess hazard model considered in this study.

2.1 Net survival

In the analysis of certain diseases survival data, the interest usually lies on analysing time since disease diagnosis until death. Since patients can died not only from the disease under study but also from other causes, when comparing disease survival between different periods of diagnosis, different regions or different socioeconomic groups for instance, it is important to have a measure that is independent from background mortality. Overall survival is thus not adequate for this type of comparison, especially in elderly patients for which other cause mortality is higher. Cause-specific survival, where only death caused by the disease in question is considered an event and all others are censored, depends on the knowledge of cause of death for all patients. In population-based data sets, this information is usually not available or is not reliable. Crude mortality quantifies the actual contribution of the disease to overall mortality. However, it is not good for comparing different regions since it also affected by background mortality [12].

Net survival is defined as the survival that would be observed in the hypothetical situation that the disease of interest is the only cause of death possible. Although this survival is not observable in the real world, it is of interest. It is the only measure that allows comparisons between different populations (originated from different regions, calendar years or other factors) since it is independent of other causes mortality [12,23]. Net survival for an individual *i* is given by the integral of the excess hazard function, i.e. the hazard due to the specific disease in study,

$$S_{N_i}(t) = exp\left(- \int_0^t \lambda_{E_i}(u) du \right)$$

The excess hazard function and modelling is described below.

2.2 Excess hazard modelling

In population-based cancer survival analysis, since cause of death is usually unknown or unreliable, the analysis is performed using relative survival methods. It is considered that the observed hazard (λ_O) can be decomposed in two additive parcels, the cancer related hazard (excess hazard) (λ_E) and the other causes hazard (λ_P), estimated by the general population mortality: λ_{Oi} (t)= λ_{Pi} (t)+ λ_{Ei}(t), for each *i* individual. The population mortality (λ_P) is obtained from life tables, usually made available by the National Statistics Offices, stratified by relevant demographic variables (e.g., sex, age, calendar year, region of residence).

The excess hazard function is modelled as a function of a set of covariates. A flexible parametric model for the excess hazard function is considered here:

$$\lambda_E(t, \boldsymbol{X}) = \lambda_0(t) \cdot \exp\big(g(\boldsymbol{t}, \boldsymbol{X})\big),$$

where $\lambda_0(t)$ is the excess hazard baseline. Following the formulation of Charvat and colleagues [24], the log of the baseline hazard is modelled using B-spline regression functions. Covariate effects expressed in $g(\boldsymbol{t}, \boldsymbol{X})$ can be parametrised with either linear or non-linear functional forms, and time-dependent effects can also be easily added in this formulation with an interaction between a B-spline of time and the covariate.

The model parameters can be estimated by maximising the full likelihood function. More details on the estimation procedure can be found in the vignette for the R package *mexhaz* by Charvat and Belot [25].

Next, we introduce one of the most common approaches to deal with missing data in statistical modelling, the multiple imputation algorithm.

2.3 Multiple imputation

Multiple imputation (MI) was first introduced by Rubin in 1978 [26]. Initially, MI was developed in the framework of survey nonresponse but has nowadays been expanded to a broader set of different fields, including survival analysis [27].

In MI, several imputations are generated for each missing value, as opposed to single imputation where each missing value is replaced by a single value. This creates several completed datasets, as many as the number of imputations performed. Each completed dataset is analysed using standard methods for complete data. The results from the several analyses are then combined to produce single estimates and confidence intervals that incorporate missing-data uncertainty.

The process can be divided in three main steps: the imputation, the analysis and the combination steps. The models related to the first step are commonly designated as imputation models and the ones used in the second step, as substantive models [28]. Briefly, the algorithm proceeds as follows:

i. Using the imputation model, generate *M>1* values for each missing value, obtaining *M* completed datasets;
ii. Fit the substantive model independently to each one of the *M* completed datasets;
iii. Combine the results obtained from each analysis performed in the previous step using Rubin's rules [29].

The MI algorithm typically relies on the assumption that the data are missing at random (MAR). This means that the probability of having a missing observation is random conditioned on the observed information, i.e. does not depend on unobserved data.

In MI the imputation phase is separated from the analysis phase. The imputation models used may thus be incompatible with the substantive model. Incompatibility means that there is no joint model for which the respective conditional distributions equal the imputation and substantive conditional models [4].

2.4 Compatibility between imputation and substantive model

To overcome the problem of incompatibility between imputation and substantive models in multiple imputation, Bartlett *et al.* [4] developed an algorithm that ensures that each covariate with missing observations is imputed from a model compatible with the substantive model. The algorithm is referred as Substantive Model Compatible-Fully Conditional Specification (SMC-FCS).

The rationale of the method is described briefly. Let *Y* represent the outcome, ***X*** a vector of *p* partially observed covariates and ***Z*** a vector of fully observed covariates. For each partially observed covariate X_j, $\boldsymbol{X}_{-j}$ represents the vector of covariates excluding that covariate ($X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$). Bartlett starts by noting that the imputation model for X_j, conditioned on the remaining covariates and the outcome is proportional to the product of the substantive model and the imputation model for X_j not involving the outcome:

$$f\left(X_j \middle| X_{-j}, Z, Y\right) = \frac{f\left(Y, X_j, X_{-j}, Z\right)}{f\left(Y, X_{-j}, Z\right)}$$

$$\propto f(Y|X, Z) \cdot f(X_j|X_{-j}, Z)$$

In the SMC-FCS algorithm, therefore, a model $f(X_j|X_{-j},Z,\phi_j)$ must be specified for each $j=1,\ldots,p$, together with noninformative priors $g(\phi_j)$. Given values of the parameters of the imputation and substantive model (ϕ_j and ψ, respectively) the missing values of X_j are imputed from a density proportional to:

$$f(Y|X, Z, \psi) \cdot f(X_j|X_{-j}, Z, \phi_j)$$

Since generally this density does not belong to a standard parametric family, drawing samples from it is non-trivial [4]. Bartlett and colleagues proposed a rejection sampling procedure that involves repeatedly drawing values for X_j from a candidate distribution,

$f(X_j|X_{-j},Z,\phi_j)$, and U from a uniform distribution on (0,1) until the drawn values satisfy the condition:

$$U \leq \frac{f(Y|X_j^*, X_{-j}, Z, \psi)}{c(Y, X_{-j}, Z, \psi)}$$

where $c(Y, X_{-j}, Z, \psi)$ is an upper bound (in X_j) for $f(Y|Xj,X\text{-}j,Z,\psi)$ that does not involve X_j.

2.5 SMC-FCS in excess hazard models

The SMC-FCS algorithm was extended here to accommodate excess hazard models. We consider that the substantive model of interest is an excess hazard model with *p* partially observed variables $\boldsymbol{X} = (X_1, \ldots, X_p)$ and a fully observed vector of *q* variables $\boldsymbol{Z}=(Z_1, \ldots, Z_q)$:

$$\lambda_E(\boldsymbol{X}, \boldsymbol{Z}, t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \lambda_0(t; \boldsymbol{\gamma}) \cdot \exp\left(g(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\beta})\right)$$

The algorithm to generate the *m*th imputed data set is as follows (adapted from [11]):

1) Using the appropriate life table (general population mortality), calculate the population hazard (λ_P) and the cumulative population hazard (Λ_P) given the demographic variables (age, sex, calendar year and region) at the time of death or

end of follow-up. Considering that the demographic variables are fully observed, this population hazard does not depend on the imputed values so it must be done only once.

2) Fill in all missing values for the incomplete variables with a starting arbitrary value (for example, mean or mode of observed values).

3) Fit the excess hazard model of interest to the current complete dataset to obtain estimates of the model parameters $(\hat{\beta}, \hat{\gamma})$ and of the respective variance-covariance matrix $\hat{\Sigma}$. Draw values $\beta^{(m)}$ and $\gamma^{(m)}$ from a joint normal distribution with means $\hat{\beta}$ and $\hat{\gamma}$ and variance-covariance matrices $\hat{\Sigma}$.

4) Calculate the estimate of the baseline excess hazard $\lambda_0^{(m)}(t)$ and of the baseline cumulative excess hazard $\Lambda_0^{(m)}(t)$ using parameter values $\gamma^{(m)}$.

5) Fit a regression model (linear, logistic, multinomial, as appropriate) of X_j on X_{-j} and Z to the current completed data set - $f(X_j|X_{-j},Z,\phi_j)$. Draw a value ϕ^* from a joint normal distribution with mean and covariance matrix given from the fitted imputation model.

6) For each individual for whom X_j is missing, (i) draw a value of X_j^* from the distribution $f(X_j|X_{-j},Z;\phi^*)$ and, (ii) draw a value U from a uniform distribution on [0,1]. Depending on the value of the censoring indicator (δ), accept the value X_j^* if:

$$U \leq exp[-\Lambda_P(t)] \cdot exp\left[-\Lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right] \qquad \text{for } \delta = 0$$

$$U \leq \frac{\left[\lambda_P(t) + \lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right] \cdot exp\left[-\Lambda_P(t) - \Lambda_0^{(m)}(t) \cdot e^{g(X_j^*,X_{-j},Z,\beta)}\right]}{\frac{\lambda_0(t)}{\Lambda_0(t)} \cdot exp\left[-\Lambda_P(t) - 1 + \frac{\Lambda_0(t) \cdot \lambda_P(t)}{\lambda_0(t)}\right]}$$

for δ = 1

Repeat (i) and (ii) until a value of X_j^* is accepted. A detailed description on the derivation of the conditions in which the rejection sampling must be done is presented in the Supplementary Material (Section S1).

7) Return to step 3 until one cycle is done for all variables with missing data.
8) Repeat steps 3-7 a certain number of iterations so that the imputed values of **X** converge to a stationary distribution. The obtained values form the *m^th^* imputed data set. Repeat the process *M* times to obtain *M* imputed datasets.

This algorithm has been implemented in R[30] software and is presented in the Supplementary Material.

3. Simulation study

A simulation study was first performed to evaluate the performance of the SMC-FCS algorithm when the substantive model of interest is an excess hazard model. This example was adapted from the one presented by Bartlett and colleagues for the Cox model [4]. Two covariates were simulated, one binary variable X_1~Be(p=0.5) and one continuous $X_2|X_1$~N(μ=X_1, σ=1). Times to death from cancer T_C were simulated from the excess hazard model: $\lambda_E(t|X) = 0.002\text{exp}\ (\beta_1 X_1 + \beta_2 X_2)$ considering $\beta_1 = \beta_2 = 1$. Times to death from other causes T_P were generated from an exponential distribution with hazard 0.001 . Censoring times C were also generated from an exponential distribution but with hazard 0.002. Finally the observed survival time was defined as T=min(T_C, T_P, C) and the event indicator δ=1 when T<C, δ=0 otherwise. Each of the 1000 simulated datasets had *n*=1000 subjects. Data on X_2 were made missing considering a MCAR mechanism such that the probability of missingness was 0.3. Missingness in X_1 was imposed considering two different scenarios: A) MCAR with probability of missingness 0.3; B) MAR dependent on outcome such that *logit* (P(X_1 miss)) = - 0.30 + 0.01T (where *T* represents survival time). In the last scenario the coefficients were chosen so that the proportion of missingness in X_1 was also around 0.3.

For each simulated dataset, three approaches for handling missing data were compared: i) Complete-case analysis (CCA), where all the cases with at least one variable missing were discarded; ii) Multiple imputation using fully conditional specification (FCS), including in the imputation models the Nelson-Aalen cumulative excess hazard estimates, the event indicator and X_1 when imputing X_2 or X_2 when imputing X_1: a logistic regression model was used for imputing X_1 and a linear regression model for X_2; iii) Multiple imputation using the substantive model compatible- fully conditional specification algorithm (SMC-FCS) as described above. Again, a logistic regression model was used for imputing X_1 using X_2 as predictor and a linear regression model for imputing X_2 using X_1 as predictor. In this algorithm, the outcomes are not included directly as covariates in the imputation models. The outcomes are included in the substantive model with which the imputed values must be compatible.

The results obtained for the two simulated scenarios are presented in Table 1 and Figures 1a) and 1b). As expected, the CCA produced unbiased estimates of the two model parameters and empirical coverages close to the nominal level of 95% when the missingness mechanism is MCAR but biased estimates when the missingness depended on the outcome (Scenario B). The conventional multiple imputation approach (FCS) produced biased estimates for both parameters and empirical coverages below 95% for both scenarios. On the contrary, the SMC-FCS algorithm produced unbiased estimates in both situations, with lower variability than CCA estimates (lower standard deviations) and with empirical coverages within the expected values.

4. Socioeconomic inequalities in survival from colorectal cancer

Colorectal cancer in the North region of Portugal

The North Region Cancer Registry of Portugal (RORENO) is a population-based cancer registry responsible for collecting information on all incidence cancer cases occurring in the North region of Portugal (~3.6 million inhabitants). The registry was set up in 1988 and in 2018 was integrated in the National Cancer Registry (RON).

A previous study [31] evaluated the existence of socioeconomic inequalities in net survival from colorectal cancer patients diagnosed in the period 2000-2002 in the area covered by RORENO. In that study, we found inequalities in net survival when using general life tables but that disappeared when including relatively small socioeconomic differences in background mortality. In the present study, we intended to update that evaluation for a more recent period, using deprivation-specific life tables recently built[32] and considering extent of disease at diagnosis as a confounder. Extent of disease is a classification defined by the European Network of Cancer Registries (ENCR) based on the TNM classification [33]. The classification is as follows: Tumour localised (T1-2N0M0); Tumour with local spread (T3-4N0M0); Tumour with regional spread (anyTN+M0); Advanced cancer (anyTanyNM1). Here, the extent was dichotomised as advanced cancer versus the other three categories (non-advanced).

More specifically, all new cancer cases of colorectal cancer (ICD10: C18-C20), diagnosed in the period 2010-2012, in patients with age at diagnosis of at least 15 years-old and below 95, residing in the North region of Portugal, were considered eligible for analysis. Only the first tumour occurring during the analysed period was considered. Second primary colorectal cancers, either synchronous or metachronous were excluded.

Survival time was considered as time between diagnosis and death from any cause or end of follow-up (31st December 2017).

Deprivation indicator

The Portuguese version of the European Deprivation Index was used as deprivation indicator. This index was built using a methodology first proposed by Pornet and colleagues in 2012[34] and then applied to five European countries: France, England, Italy, Spain and Portugal [35]. The index is based on census variables available for each country that are most associated with variables identified from the European Union Statistics on Income and Living Conditions (EU-SILC) survey [36]. The index for Portugal based on 2001 census includes percentage of: non-owned households, households without indoor flushing, residents with low education level (≤6th grade), household with 5 rooms or less, unemployed looking for a job, female residents aged 65 years or more, households without bath/shower and percentage of residents employed in manual occupations [37]. A score was obtained for each parish based on the census responses of its inhabitants. This score was then categorized in five quintiles from the least deprived (q1) to the most deprived (q5) such that each quintile corresponded to 20% of the Portuguese population. Each patient was assigned with the deprivation quintile corresponding to his/her parish of residence at the time of diagnosis.

Data description

A total of 8108 new cancer cases were considered eligible for analysis. After excluding patients with unknown status at the end of follow-up (n=154; 1.9%), a total of 7954 patients were included in the analysis. Distribution of cases by age group, cancer site, deprivation quintile and extent of disease at diagnosis was calculated by sex (Table 2).

Male patients represented 58.6% of the cohort. Women presented a higher median age compared to men: 71 vs 69 years (p<0.001). The proportion of rectum cancer cases was higher in men (p=0.035). No differences were found in the distribution by deprivation groups between male and female patients (p=0.208). Also, the distribution of extent of disease at diagnosis was similar between both sexes (p=0.206).

Missing data

A very low proportion of cases had missing information on deprivation quintile (0.5%). We thus decided to exclude these cases from further analysis. Extent of disease at diagnosis is the main prognostic variable and had a considerable proportion of missing data (40.3%). To evaluate which variables were associated with missingness in extent of disease, a multivariable logistic regression model was built considering missing extent as the outcome. Variables included in the model were sex, age group, tumour site (colon vs rectum), EDI deprivation quintile, basis of diagnosis, vital status at the end of follow-up and survival time in years. Sex was not associated with extent missingness. Older patients and patients without a microscopically verified diagnosis had increased odds of having unknown extent. Rectum cancer patients and patients living in more deprived areas had lower odds of extent missingness. Survival time and vital status were also associated with the chances of having missing extent (Table 3).

Age-standardised net survival (ASNS) at 1-year of the patients with known extent (84.2%; 95%CI: 83.1-85.2) was higher than ASNS of patients with missing extent information (80.7%; 95%CI: 79.4-82.1). On the contrary, ASNS at 5-years was higher in patients with unknown extent (67.1%; 95%CI: 65.2-69.1) than with known extent (63.9%; 95%CI: 62.3-65.6).

Results

The main aim of the analysis performed was to evaluate the existence of socioeconomic inequalities in net survival from colorectal cancer in the cohort of patients described above, while considering the following potential confounders: age, sex and extent of disease at diagnosis. The proportion of cases with missing extent was around 40%.

First, net survival by SE group was estimated for the full dataset using the non-parametric Pohar-Perme estimator [23]. Differences between net survival curves were assessed using the log-rank-type test developed by Grafféo and colleagues [38].

The unadjusted net survival curves (Figure 2) showed a better net survival for patients living in least deprived areas (p=0.010). Five-year net survival was 66.9% for the least deprived group and 62.0% for the most deprived one.

Second, excess hazard ratios were estimated. Missing data was handled using complete-case analysis and multiple imputation using the standard FCS and the adapted SMC-FCS approach. Covariates considered in the substantive model were age, deprivation index (EDI), sex and extent of disease at diagnosis. All covariates were assumed to have no time-dependent effects. The excess hazard baseline was modelled using B-splines with one knot at one year of follow-up.

In this example, only one covariate had missing data (extent). The imputation model in the standard FCS approach included as covariates age, sex, EDI, tumour site and basis of diagnosis besides the event indicator and the cumulative excess hazard estimated by the Nelson-Aalen estimator. In the SMC-FCS approach, the same variables were used in the imputation model except the "outcome", namely the cumulative excess hazard baseline and the event indicator. In this approach, the outcomes are indirectly accounted for in the rejection sampling algorithm which guarantees compatibility between the imputation and substantive models. In both MI approaches, extent of disease was imputed using a binomial logistic regression model. Fifty imputations were used in each approach.

The results obtained for the excess hazard ratios (EHR) using the three different approaches are presented in Table 4. The estimated EHRs using the complete-case analysis and the FCS approach were similar. Using SMC-FCS, there was attenuation on the excess hazard ratio of advanced tumours vs non-advanced.

Using the SMC-FCS algorithm, there was a more clear trend in the EHRs by deprivation quintile showing an increased excess hazard for patients coming from the more deprived areas (although not reaching statistical significance).

5. Discussion

The SMC-FCS approach to MI was first proposed by Bartlett and colleagues to ensure compatibility of the imputation models with the substantive model [4]. The algorithm relies on a rejection sampling scheme. The conditions of acceptance of a proposed imputation value depend on the substantive model of interest. These conditions were derived in this study for the situation where the substantive model is an excess hazard model. This type of model is very common in population-based cancer survival analysis while missing data in population-based cancer research are also common. The algorithm for binary and continuous covariates was implemented in R and is presented on the Appendix.

The proposed adaptation of the SMC-FCS algorithm to cope with excess hazard models was tested in a simulation study for two different scenarios of missingness. When missingness was MCAR, the complete-case analysis produced unbiased estimates as expected. In the second scenario, where missingness was dependent on the outcome (survival time), the model parameter estimates obtained in the complete-case analysis were biased, including the parameter of the variable for which the missingness mechanism was MCAR. The standard FCS multiple imputation approach produced biased estimates and poor empirical coverages for both parameters. These results were observed in both missingness scenarios analysed. Due to the non-linear nature of the substantive model considered (excess hazard model), the FCS approach does not guarantee the compatibility between the imputation and substantive models. On the contrary, the SMC-FCS approach to MI produced unbiased estimates of both parameters in all scenarios. Also, the standard errors of the estimates were lower than for the complete-case analysis. These results confirm that also when the substantive model is an excess hazard model, the SMC-FCS approach presented lower bias and coverage closer to the nominal value of 95% relatively to the other two approaches.

In the example analysed, missing extent was showed to be associated with vital status and survival time. These are the outcomes of interest in survival analysis and excess hazard modelling. This shows that using a complete case analysis would result most certainly in biased results. Also, the net survival probability of patients with known extent of disease was significantly different from the one of patients with unknown extent, therefore not favouring the hypothesis of the extent being missing completely at random.

One of the advantages associated with multiple imputation is the possibility of using variables in the imputation model that are not of interest in the substantive model, to increase the plausibility of the MAR assumption and the efficiency of the imputation process. In the SMC-FCS algorithm, to draw imputations that are compatible with the substantive model the variables considered in both imputation and substantive models during the imputation process must be the same. It is possible, however, to fit models to the imputed datasets in which fewer explanatory variables are used [4]. In the example analysed, two auxiliary variables were used in the imputation process (basis of diagnosis and tumour site) since these have been shown to be related with the chance of extent being missing.

No major differences in the estimated adjusted effects of socioeconomic condition on the excess hazard were observed between the CCA and the classical MI approach. However, when using SMC-FCS, the trend for higher EHRs in the more deprived areas was clearer.

In MI the missing values are imputed using imputation models dependent on a set of covariates. The efficiency of these imputations depends on the availability of variables that are both associated with the probability of missingness and with the missing variable. In this study, the number of variables used in the imputation model was low and their association with extent of disease was weak, which may have diminished the efficiency of the imputations performed.

We have implemented SMC-FCS for excess hazard models considering binary and continuous covariates. Work is in progress to extend the algorithm to categorical covariates with more than 2 categories.

In this study, the proportional hazards assumption was assumed for all variables. We acknowledge that the effect of some covariates can typically be time-dependent. A first approach for extending the SMC-FCS approach to cope with excess hazard models was presented. Further research is needed to include time-dependent effects in excess hazard models following the work that Keogh and Morris have done for the Cox models [11].

Competing interests

The authors declare that they have no competing interest.

Ethics approval and consent to participate

This study was approved by the Ethical Committee of the Portuguese Oncology Institute of Porto, Portugal (Refª 203/018).

References

1. Murray JS. Multiple Imputation: A Review of Practical and Theoretical Findings. *Stat Sci* 2018; 33: 142–159.
2. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Med Res Methodol* 2015; 15: 30.
3. Yucel RM. State of the Multiple Imputation Software. *J Stat Softw*; 45.
4. Bartlett JW, Seaman SR, White IR, et al. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat Methods Med Res* 2015; 24: 462–487.
5. Barzi F, Woodward M. Imputations of Missing Values in Practice: Results from

Imputations of Serum Cholesterol in 28 Cohort Studies. *Am J Epidemiol* 2004; 160: 34–45.

6. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data. *J Clin Epidemiol* 2003; 56: 28–37.

7. Marshall A, Altman DG, Royston P, et al. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 2010; 10: 7.

8. van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999; 18: 681–94.

9. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009; 28: 1982–98.

10. Bartlett JW, Morris TP. Multiple imputation of covariates by substantive-model compatible fully conditional specification. *Stata J* 2015; 15: 437-456(20).

11. Keogh RH, Morris TP. Multiple imputation in Cox regression when there are time-varying effects of covariates. *Stat Med* 2018; 37: 3661–3678.

12. Pohar Perme M, Estève J, Rachet B. Analysing population-based cancer survival – settling the controversies. *BMC Cancer* 2016; 16: 933.

13. Belot A, Ndiaye A, Luque-Fernandez M-A, et al. Summarizing and communicating on survival data according to the audience: a tutorial on

different measures illustrated with population-based cancer registry data. *Clin Epidemiol* 2019; Volume 11: 53–65.

14. Uhry Z, Bossard N, Remontet L, et al. New insights into survival trend analyses in cancer population-based studies. *Eur J Cancer Prev* 2017; 26: S9–S15.

15. Belot A, Remontet L, Rachet B, et al. Describing the association between socioeconomic inequalities and cancer survival: methodological guidelines and illustration with population-based data. *Clin Epidemiol* 2018; 10: 561–573.

16. Giorgi R, Belot A, Gaudart J, et al. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Stat Med* 2008; 27: 6310–6331.

17. Nur U, Shack LG, Rachet B, et al. Modelling relative survival in the presence of incomplete data: a tutorial. *Int J Epidemiol* 2010; 39: 118–128.

18. Dejardin O, Rachet B, Morris E, et al. Management of colorectal cancer explains differences in 1-year relative survival between France and England for patients diagnosed 1997-2004. *Br J Cancer* 2013; 108: 775–83.

19. Dejardin O, Jones AP, Rachet B, et al. The influence of geographical access to health care and material deprivation on colorectal cancer survival: Evidence from France and England. *Heal Place* 2014; 30: 36–44.

20. Walters S, Maringe C, Butler J, et al. Breast cancer survival and stage at diagnosis

in Australia, Canada, Denmark, Norway, Sweden and the UK, 2000-2007: A population-based study. *Br J Cancer* 2013; 108: 1195–1208.

21. Le Guyader-Peyrou S, Orazio S, Dejardin O, et al. Factors related to the relative survival of patients with diffuse large B-cell lymphoma in a population-based study in France: does socio-economic status have a role? *Haematologica* 2017; 102: 584–592.

22. Falcaro M, Nur U, Rachet B, et al. Estimating Excess Hazard Ratios and Net Survival When Covariate Data Are Missing. *Epidemiology* 2015; 26: 421–428.

23. Perme MP, Stare J, Estève J. On Estimation in Relative Survival. *Biometrics* 2012; 68: 113–120.

24. Charvat H, Remontet L, Bossard N, et al. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Stat Med* 2016; 35: 3066–84.

25. Charvat H, Belot A. *Analysis of time-to-event data with mexhaz*. 2020.

26. Rubin DB. Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. 1978, pp. 20–34.

27. Carpenter JR, Kenward MG. *Multiple imputation and its application*. John Wiley & Sons, 2013.

28. Carpenter JR, Kenward MG. *Missing data in randomised controlled trials a practical guide*. Birmingham: Health Technology Assessment Methodology Programme, 2007.

29. Rubin DB, Wiley InterScience. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

30. R Core Team. R: A Language and Environment for Statistical Computing.

31. Antunes L, Mendonça D, Bento MJ, et al. No inequalities in survival from colorectal cancer by education and socioeconomic deprivation - a population-based study in the North Region of Portugal, 2000-2002. *BMC Cancer* 2016; 16: 608.

32. Antunes L, Mendonça D, Ribeiro AI, et al. Deprivation-specific life tables using multivariable flexible modelling – trends from 2000–2002 to 2010–2012, Portugal. *BMC Public Health* 2019; 19: 276.

33. Berrino F, Brown C, Moller T, et al. *ENCR RECOMMENDATIONS, Condensed TNM for Coding the Extent of Disease*. Lyon, 2002.

34. Pornet C, Delpierre C, Dejardin O, et al. Construction of an adaptable European transnational ecological deprivation index: the French version. *J Epidemiol Community Health* 2012; 66: 982–9.

35. Guillaume E, Pornet C, Dejardin O, et al. Development of a cross-cultural

deprivation index in five European countries. *J Epidemiol Community Health* 2015; jech-2015-205729.

36. Eurostat. Access to Microdata. European Union Statistics on Income and Living Conditions (EU-SILC), http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions (2015, accessed 19 May 2018).

37. Ribeiro AI, Mayer A, Miranda A, et al. *Acta Médica Portuguesa.* 2017.

38. Grafféo N, Castell F, Belot A, et al. A log-rank-type test to compare net survival distributions. *Biometrics* 2016; 72: 760–9.

Table 1 – Comparison of excess hazard models parameters estimates for different approaches of missing data handling. Results from n=1000 simulations.

	CCA			**FCS**			**SMC-FCS**		
	Mean	SD	Cov	Mean	SD	Cov	Mean	SD	Cov
Scenario A									
$\beta_1 = 1$	1.001	0.143	*95.2*	0.929	0.124	*93.4*	1.003	0.126	*95.6*
$\beta_2 = 1$	1.004	0.069	*95.7*	0.858	0.053	*50.7*	1.004	0.057	*95.8*
Scenario B									
$\beta_1 = 1$	0.855	0.128	*89.4*	0.895	0.128	*89.5*	1.008	0.128	*95.4*
$\beta_2 = 1$	0.819	0.068	*44.7*	0.880	0.051	*62.5*	1.001	0.058	*95.5*

Scenario A: X1, X2 MCAR

Scenario B: X1 MAR dependent of outcome, X2 MCAR

CCA – Complete-case analysis

FCS – Fully conditional specification

SMC-FCS – Substantive model compatible FCS

Table 2 - Sociodemographic and clinical characteristics of the colorectal cancer patients (2010-2012).

Variable	Male		Female		Total	
	n	%	n	%	n	%
Total by sex	4 664	58.6	3 290	41.4	7 954	100.0
Age group						
15-44	177	3.8	153	4.7	330	4.1
45-54	460	9.9	334	10.2	794	10.0
55-64	1 072	23.0	662	20.1	1 734	21.8
65-74	1 415	30.3	845	25.7	2 260	28.4
75+	1 540	33.0	1 296	39.4	2 836	35.7
Tumour site						
Colon	3 060	65.6	2 234	67.9	5 294	66.6
Rectum	1 604	34.4	1 056	32.1	2 660	33.4
Deprivation (EDI)						
q1 (least deprived)	444	9.5	337	10.2	781	9.8
q2	609	13.1	415	12.6	1 024	12.9
q3	1 074	23.0	693	21.1	1 767	22.2
q4	1 233	26.4	894	27.2	2 127	26.7
q5 (most deprived)	1 280	27.4	939	28.5	2 219	27.9
Unknown	24	0.5	12	0.4	36	0.5
Tumour extent at diagnosis						
Non-advanced	2 147	46.0	1 454	44.2	3 601	45.3
Advanced	636	13.6	502	15.3	1 138	14.3
Unknown	1 881	40.3	1 334	40.5	3 215	40.4

Table 3 - Sociodemographic characteristics of patients with known extent vs patients with unknown extent[a]. Odds ratio of having missing extent.

	Extent of disease at diagnosis					
Variable	**Known**		**Unknown**		**Unknown vs known**	
	n	**%**	**n**	**%**	**OR[b]**	**95%CI**
Total by extent	4 725	59.7	3 193	40.3		
Sex						
Male	2 771	58.6	1 869	58.5	1	
Female	1 954	41.4	1 324	41.5	0.95	0.86 - 1.04
Age group						
15-44	224	4.7	103	3.2	1	
45-54	510	10.8	279	8.7	1.19	0.90 - 1.58
55-64	1 085	23.0	645	20.2	1.25	0.97 - 1.62
65-74	1 359	28.8	891	27.9	1.37	1.07 - 1.77
75+	1 547	32.7	1275	39.9	1.73	1.35 - 2.23
Tumour site						
Colon	2 959	62.6	2312	72.4	1	
Rectum	1 766	37.4	881	27.6	0.65	0.59 - 0.72
Deprivation (EDI)						
q1 (least deprived)	430	9.1	351	11.0	1	
q2	631	13.4	393	12.3	0.74	0.61 - 0.90
q3	1 055	22.3	712	22.3	0.81	0.68 - 0.97
q4	1 258	26.6	869	27.2	0.83	0.70 - 0.99
q5 (most deprived)	1 351	28.6	868	27.2	0.77	0.65 - 0.92
Basis of diagnosis						
Microscopically verified	4 614	97.7	2 904	90.9	1	
Non-micros. Verified	111	2.3	289	9.1	4.00	3.19 - 5.05
Vital status at end of follow-up						
Alive	2 440	51.6	1 674	52.4	1	
Dead	2 285	48.4	1 519	47.6	0.63	0.53 - 0.75
Survival time						
Mean	4.22	-	4.07	-	0.94	0.91 – 0.97

a) Cases with unknown deprivation EDI (n=36; 0.5%) were not considered in this analysis.

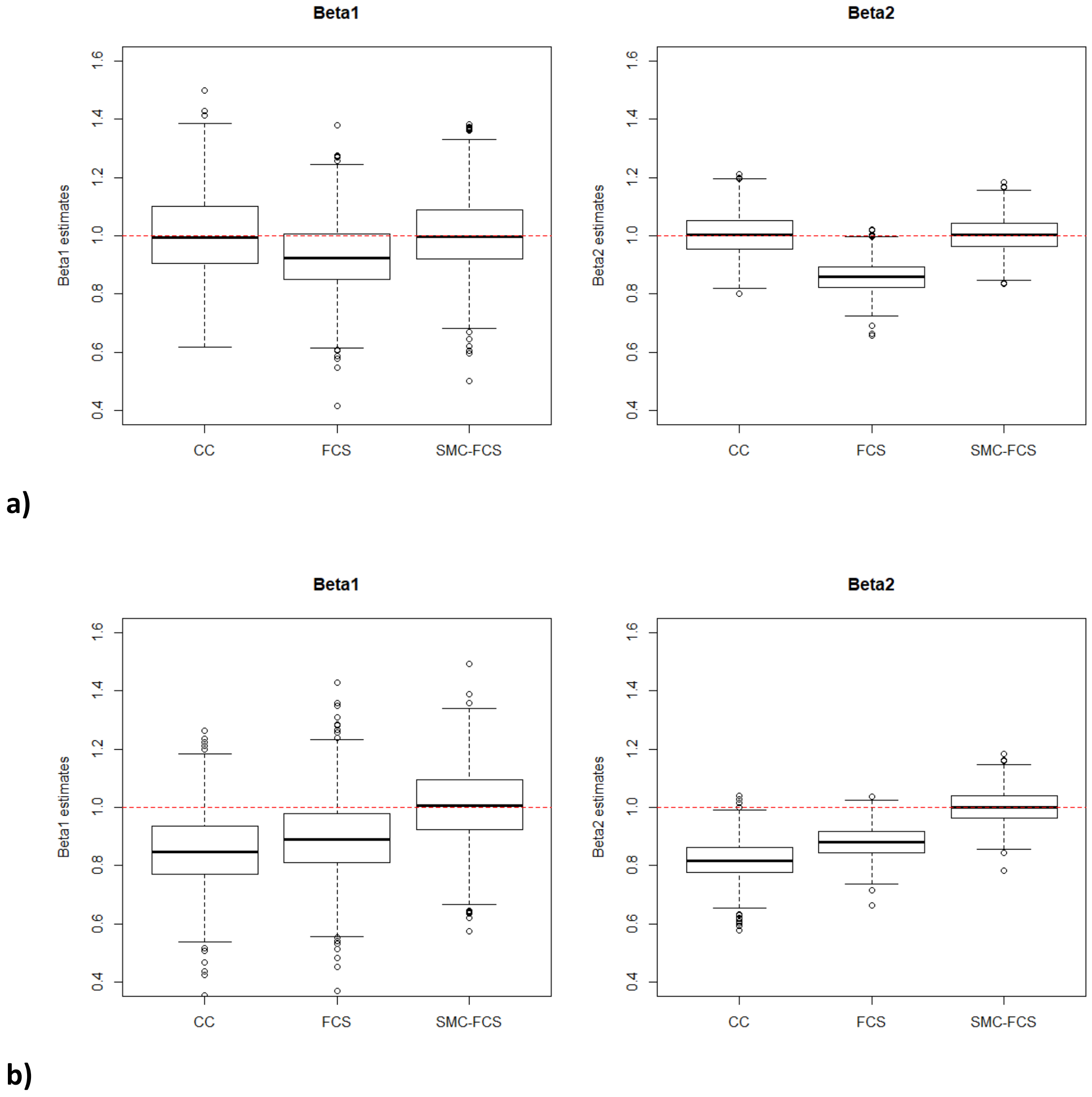
b) Adjusted ORs (adjusted for sex, age group, tumour site, deprivation, basis of diagnosis, vital status and/or survival time)

Table 4 - Excess hazard ratios (CCA; FCS MI; SMC-FCS MI)

Variable	CCA		FCS MI		SMC-FCS MI	
	EHR[a)]	95%CI	EHR[a)]	95%CI	EHR[a)]	95%CI
EDI						
q1	1		1		1	
q2	1.02	0.81 - 1.28	1.01	0.82 - 1.23	1.05	0.82 - 1.34
q3	1.04	0.84 - 1.28	1.10	0.91 - 1.33	1.15	0.94 - 1.40
q4	1.13	0.93 - 1.38	1.09	0.91 - 1.30	1.14	0.93 - 1.40
q5	1.08	0.89 - 1.33	1.16	0.97 - 1.39	1.20	0.99 - 1.46
Extent						
Non-advanced	1		1		1	
Advanced	10.1	9.02 - 11.3	10.0	8.88 - 11.2	8.35	6.81 - 10.2

a) Adjusted for age, sex and EDI or extent.

Figure 1 - Comparison of excess hazard models parameters estimates for different approaches of missing data handling. Results from n=1000 simulations (a – Scenario A: MCAR; b – Scenario B: MAR)



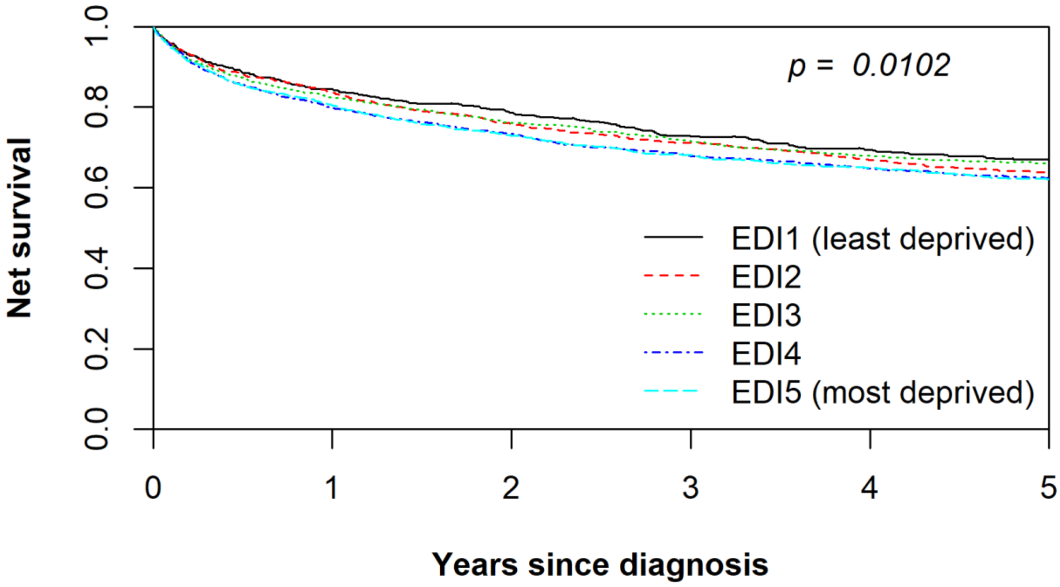


Figure 2 – Net survival by EDI category for the full cohort.