Published in final edited form as:
Biom J. 2019 July ; 61(4): 1020–1032. doi:10.1002/bimj.201800275.

Cox regression model with randomly censored covariates

Folefac D. Atem[1], **Roland A. Matsouaka**[2,3], **Vincent E. Zimmern**[4,5]

[1]Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, Houston, TX, USA

[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[3]Program for Comparative Effectiveness Methodology, Duke Clinical Research Institute, Duke University, Durham, NC, USA

[4]Department of Pediatrics, University of Texas Southwestern Medical School, Dallas, TX, USA

[5]Department of Pediatrics, Children Hospital Dallas, Dallas, TX, USA

Abstract

This paper deals with a Cox proportional hazards regression model, where some covariates of interest are randomly right-censored. While methods for censored outcomes have become ubiquitous in the literature, methods for censored covariates have thus far received little attention and, for the most part, dealt with the issue of limit-of-detection. For randomly censored covariates, an often-used method is the inefficient complete-case analysis (CCA) which consists in deleting censored observations in the data analysis. When censoring is not completely independent, the CCA leads to biased and spurious results. Methods for missing covariate data, including type I and type II covariate censoring as well as limit-of-detection do not readily apply due to the fundamentally different nature of randomly censored covariates. We develop a novel method for censored covariates using a conditional mean imputation based on either Kaplan–Meier estimates or a Cox proportional hazards model to estimate the effects of these covariates on a time-to-event outcome. We evaluate the performance of the proposed method through simulation studies and show that it provides good bias reduction and statistical efficiency. Finally, we illustrate the method using data from the Framingham Heart Study to assess the relationship between offspring and parental age of onset of cardiovascular events.

Keywords

censored covariate; complete-case analysis; Cox proportional hazards model; random censoring; survival analysis

Correspondence: Folefac D. Atem, University of Texas School of Public Health, 6011 Harry Hines V8.112, Dallas, TX 75390-9128, USA. folefac.d.atem@uth.tmc.edu.

CONFLICTS OF INTEREST
The authors declare that there is no conflict of interest.

SUPPORTING INFORMATION
Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

1 | INTRODUCTION

Time-to-event data are ubiquitous in biomedical, social, behavioral, and epidemiological studies. Nevertheless, the issue of randomly censored covariates in Cox proportional regression model has been largely overlooked compared to censored outcomes, especially in the studies of time-to-event outcomes where censoring may be due to loss of follow-up, drop out, study termination or detection limits (Klein & Moeschberger, 2005; Kalbfleisch & Prentice, 2011). For instance, randomly-censored covariates arise in the setting of cardiovascular disease analyses. An important question in the field of cardiovascular epidemiology is how best to quantify associations between parental risk factors and the onset of cardiovascular disease in their offspring. Often, parental risk factors, such as age-of-onset of disease in parents, are randomly right-censored, meaning that the study either terminates prior to a cardiovascular event being observed or a patient is lost to follow-up prior to a cardiovascular event. In these cases, for a sensible data analysis, the goal is to impute a value for censored age-of-onset of disease in the parents, under some conditions. This allows us to include those patients, eliminate completely or at least reduce bias, and power the study appropriately. In this paper, we illustrate a new method for imputation of these randomly censored covariates using the Framingham Heart Study—an ongoing study of cardiovascular disease risk over two generations. We show that this method can be used to analyze the relationship between randomly censored covariates (the parental age of onset of cardiovascular disease) and a time-to-event outcome variable (the age of onset of cardiovascular disease in their children).

Until recently, the issue of censored covariates was addressed only in the context of type I and type II, or limit-of-detection censoring (Bernhardt, Wang, & Zhang, 2014; Helsel et al., 2005; Helsel, 2006, 2011; Sattar, Sinha, Wang, & Li, 2015; Wu, Chen, Ware, & Koyama, 2012; Yue & Wang, 2016). However, it should be borne in mind that these types of censoring are different from random censoring. In type I censoring, participants in a study are followed for a prespecified, fixed duration of time, whereas in type II censoring, the study is stopped when a prespecified number of participants are observed to experience the event of interest. Limit-of-detection censoring concerns censoring that occurs in measurement systems or instruments. Instead of follow-up times, the focus is on the prespecified level(s) of detection under (or over which) the precision of the system or the instrument is not sufficient enough to discern, measure, and quantify to an acceptable level of accuracy certain results of an experiment. When such a result is not accurately measured, it is recorded as being below (or above) a given threshold value and thus considered as censored (Cole, Chu, Nie, & Schisterman, 2009; Helsel et al., 2005; Helsel, 2011; Kong & Nan, 2016; Lee et al., 2017). In addition, censoring due to limits of detection can be characterized as type I (if the measurement lies below or above a prespecified threshold) or type II censoring (when an experiment is carried out until it reaches a prespecified number of detected measurements).

Unlike type I and type II time censoring, random censoring occurs where study participants are censored at varying time points (Kalbfleisch & Prentice, 2011). In the same way that methods for outcomes that are subject to type I, type II, or limit-of-detection censoring do not apply to randomly censored outcomes, methods related to these types of censoring may not work well with randomly censored covariates.

Furthermore, data with randomly-censored covariates are often mistaken for missing covariate data and analyzed as such. While missing and censored covariate observations share some similarities, they are fundamentally different. An observation is missing when its measurement is unavailable, either by design or by accident. However, a censored observation is an observation for which the true information is only partially observed due to a wide range of reasons, including loss due to follow-up, drop out, study termination, end of the study, or even an inherent limit-of-detection of an instrument's measurements. Although there are several well-established methods to handle missing covariates, they cannot be used for censored covariates as they do not fully capture the essence of these covariates and cannot use all possible (partial) information contained in the observations with censored covariates.

For this paper, we consider a survival analysis using Cox proportional hazards model, where some values of the covariates of interest are randomly censored. Such censored covariate observations arise as a result of a time lag between the time of measurement of the covariate (usually at baseline) and the occurrence of an event of particular interest that is closely related to the covariate and assured its availability (Atem, Sampene, & Greene, 2017; Lee, Park, & Park, 2003; Tsimikas, Bantis, & Georgiou, 2012). Suppose, for instance, that we wish to study the impact of early onset of cardiovascular events in parents as a predictor of age-of-onset of cardiovascular disease in children. Since not all parents or offspring may have experienced the event of interest by the end of the study, some observations are likely to be randomly censored.

A common method that is used to analyze data with censored covariates is the complete-case analysis where observations with censored covariate measures are simply deleted from the data analysis. While the complete-case analysis may provide unbiased estimates of the regression coefficients under specific conditions (e.g., independent censoring or small percentage of censored observations), removing censored observations can drastically reduce the sample size and, thereby, render this approach highly inefficient, especially when the percentage of censoring is high (Lipsitz, Parzen, Natarajan, Ibrahim, & Fitzmaurice, 2004). This is particularly crucial when there are multiple censored covariates since the total percentage of censored observations in the data set can be substantial. Using closely related techniques such as the available-case analysis may lead to the same issues (Rigobon & Stoker, 2009).

Moreover, it has been demonstrated that using ad hoc imputation methods where, for instance, censored observations are replaced by a constant value, the covariate mean, or the median will result in biased estimates of parameters (Atem, Sampene, & Greene, 2017; Bernhardt, Wang, & Zhang, 2014; Sattar, Sinha, & Morris, 2012; Sattar, Sinha, Wang, & Li, 2015). Finally, while dichotomizing continuous covariates is already a bad idea (Fitzsimons, 2008; Royston, Altman, & Sauerbrei, 2006), it is even worse when such a variable is censored (Austin & Hoch, 2004; Rigobon & Stoker, 2009).

More elaborate methods based on maximum likelihood estimation of fully parametric models (Langohr, Gómez, & Muga, 2004; Sattar et al., 2012), or on Cox proportional hazards regression model have been proposed (D'Angelo & Weissfeld, 2008; Chen, Wu, Ware, & Koyama, 2014; Dinse et al., 2014). Other methods such as semiparametric methods (Kong & Nan, 2016), multiple imputation procedures for accelerated failure time models (Bernhardt et al., 2014), or on M-regression models (Wang & Feng, 2012) have been developed and shown to work well for type I, type II, or limit-of-detection censoring, under some specific conditions. However, these methods cannot be readily applied to censored covariates in survival analysis where the focus is on random censoring, which is a totally different type of censoring (Atem et al., 2017).

While there have been extensive studies for other types of censored covariates, a handful of papers have looked into this issue of randomly censored covariates in classical linear and logistic regressions as well as other generalized linear models (Atem et al., 2017; Lee et al., 2003; Tsimikas et al., 2012). Methods for handling survival analysis in the presence of censored covariates have received less attention. Presently, randomly censored covariates are gaining ground with recent publications by Lee et al. (2003); Atem, Qian, Maye, Johnson, and Betensky (2017); Atem and Matsouaka (2017); Atem, Qian, Maye, Johnson, and Betensky (2016).

In this paper, we develop a non-parametric and semi-parametric conditional imputation method for a right-censored covariate. Unlike Lee's method, which is empirically based on the risk function and partial likelihood estimation (Lee et al., 2003), our proposed method replaces censored observations with conditional mean values predicted either by a non-parametric (Kaplan–Meier) or a semi-parametric estimation (Cox proportional hazards model) technique, while maximizing the use of available information in the data. This technique is an improvement to the conditional imputation of Atem et al. for linear regression (Atem et al., 2017); it can be easily programmed and implemented using common statistical software packages.

The remainder of this paper is organized as follows. First, we introduce an illustrating example based on the Framingham Heart Study. Then, in Section 3, we develop a general framework for the proposed method. In Section 4, we present simulation studies where we compare the complete-case and the proposed conditional imputation method. In Section 5, we apply our method to the study of association between offspring age at cardiovascular event and parent age of onset of cardiovascular event. Finally, we end the paper with a discussion in Section 6.

2 | ILLUSTRATING EXAMPLE: THE FRAMINGHAM HEART STUDY

The Framingham Heart Study is an ongoing prospective study of the etiology of cardiovascular diseases (CVDs). The study began in 1948 and enrolled 5,209 men and women aged between 28 and 62 years as part of the original cohort. In 1971, the Offspring Study began with a sample of 5,124 men and women aged 5–70 years who were either (genetic or adoptive) offspring or spouses of offspring of the original cohort. Approximately every 4 years, study participants are examined to update their health status information and potential risk factors. Standard clinical examination includes physician interview, physical examination, and laboratory tests. Over the years, close monitoring of Framingham Study participants has been crucial in identifying major CVD risk factors (Chen & Levy, 2016)

In this paper, we considered the available data of 2,622 participants from the original cohort (Exam 12; years 1971–1974) and 2622 participants from the offspring cohort (Exam 1; years 1971–1975). The Framingham Heart Study reviews and adjudicates events for the occurrence of cardiovascular diseases. When available, parental age of cardiovascular event and offspring age of cardiovascular event are recorded. In our data set, a total of 1,816 participants from the original cohort and 671 of the offspring cohort had a cardiovascular event.

We are interested in estimating the parameters of a Cox proportional hazards model of the relationship between age of parent’s onset of a clinically diagnosed cardiovascular event and the time-to-onset of cardiovascular disease on offspring participants, looking at both the Original and Offspring cohorts. We used age as a time scale to run the survival analysis of the time-to-cardiovascular event in offspring.

3 | PROPOSED METHOD

3.1 | Notations and definitions

Suppose N study participants are independently sampled from a reference population. For each participant i, let T_i and C_i be the true survival and censoring times respectively, $\mathbf{X}_i$ the K-vector of potentially censored covariates, R_i the covariate-censoring time, and $\mathbf{Z}_i$ the M-vector of fully observed covariates. Due to the right-censoring in survival times T and the covariate X, the data observed are $\{(Y_i, \Delta_i, V_{1i}, D_{1i}, \dots, V_{Ki}, D_{Ki}, \mathbf{Z}_i)^{\mathrm{T}} : i = 1, 2, \dots, N\}$, where $V_{ki} = \min(X_{ki}, R_{ki})$, $D_{ki} = I(X_{ki} \leq R_{ki})$, $Y_i = \min(T_i, C_i)$, and $\Delta_i = I(T_i \leq C_i)$, for $K = 1, \dots, K$.

For simplicity, we start with one censored covariate X, that is, $K = 1$ and seamlessly generalize the methodology to more than one censored covariate later. To formalize our method, we are interested in assessing the relationship between the potentially randomly right censored covariate X and the survival time, adjusting for the effects of fully observed covariates $\mathbf{Z}_i$ through a Cox proportional hazards model

$$h(t) = h_o(t)\exp\left(\beta X + \gamma^{\top}\mathbf{Z}\right), \tag{1}$$

where $h_o(t)$ is the baseline hazard, β and γ^{T} are the unknown regression coefficients we need to estimate.

Throughout this paper, model 1 is our substantive model, which we assume is correctly specified. We also assume that both R and C are random right-censoring mechanisms and non-informative, that is, $C \perp T(X, \mathbf{Z})$, $C \perp X/\mathbf{Z}$ and $R \perp X/(T,\mathbf{Z})$. The last assumption $R \perp X|(T, \mathbf{Z})$ implies that the censoring in X may depend on the fully measured covariates $\mathbf{Z}$ as well as on the time-to-event outcome (Rathouz, 2007). It can be relaxed if there is a set of auxiliary variables, that is, variables that are not inherently of interest and do not provide additional information on the hazard model of T beyond what can be obtained with (X, **Z**) alone, but may potentially contain valuable information to predict censoring R and thus enhance the imputation of censored observations (Collins, Schafer, & Kam, 2001; Zhou & Pepe, 1995). Finally, we denote by f the density distribution of X (or that of (V, D) depending on the context) and by f_T and F_T, respectively, the conditional density and cumulative distribution functions for the survival time T. Similarly, we also used the notations f_C and F_C for the censoring mechanism C.

3.2 | Imputation procedure

An imputation model for the covariate X can be obtained by specifying the conditional distribution $f(V, D/\mathbf{Z})$. However, if such an imputation model is not compatible with the substantive model, the imputation procedure may lead to specious results. As suggested by Bartlett et al. (2015), such an incompatibility can be avoided if there is a joint model for the outcome and the covariate of interest from which we deduce an imputation model or algorithm. Our imputation model is similar to the method proposed by Rubin (2004), Schafer (1999), and Meng (1994). In order to eliminate inconsistency, they proposed that the assumptions in both models (imputer and analyst model) should be similar and the imputer model should not make more assumptions than the analyst model. The conditional distribution of such a joint model, given the available covariates, would correspond to the given (correctly) specified substantive model.

In the absence of covariate censoring, we can use the factorization,

$$f(X \mid T, \mathbf{Z}) = \frac{f(T, X \mid \mathbf{Z})}{f(T \mid \mathbf{Z})} = \frac{f(T \mid X, \mathbf{Z}) \times f(X \mid \mathbf{Z})}{f(T \mid \mathbf{Z})},$$

which implies that $f(X/T,\mathbf{Z}) \propto f(X/\mathbf{Z})f(T/X, \mathbf{Z})$.

When covariate censoring is present, we observe both (V, D) for the covariate measure and (Y, Δ) for the outcome. The above factorization can be used to write

$$f(V, D \mid Y, \Delta, \mathbf{Z}) \propto f(V, D \mid \mathbf{Z}) f(Y, \Delta \mid X, \mathbf{Z}) . \tag{2}$$

This allows us to elucidate the second assumption in Section 3.1 and demonstrate the role Y plays in imputing censored values of X.

A, case in point, when T is censored at time $C = c$, the joint density $f(V, D/Y, \Delta, \mathbf{Z})$ is proportional to $S_T(c/V,D,\mathbf{Z})f_c(c/V,D,\mathbf{Z})f((V,D)|\mathbf{Z})$. On the other hand, when the survival time T is observed at time $T = t$, the right-hand side of (3) becomes $f_T(t/V,D,\mathbf{Z})[1 - F_C(t/V,D,\mathbf{Z})]f((V,D)|\mathbf{Z})$**.** With assumption $C \perp X/\mathbf{Z}$, the factorization simplifies to

$$f(V, D \mid Y, \Delta, \mathbf{Z}) \propto \left[\left(1 - F_T(c \mid \mathbf{Z})\right) f_C(c \mid \mathbf{Z})\right]^{1 - \Delta}\left[f_T(t \mid V, D, \mathbf{Z})\left[1 - F_C(t \mid \mathbf{Z})\right]\right]^{\Delta} f((V, D) \mid \mathbf{Z}) . \tag{3}$$

Therefore, to obtain a genuine and compatible imputation of X (also known as congenial imputation (Meng, 1994)) using the observed data, the left-hand side of the above factorization (2) requires that leveraging the information provided by the fully observed covariates Z, observed measures of X as well as Y, while accounting for the role of the censoring mechanisms D and Δ (Meng, 1994; Collins, Schafer, & Kam, 2001; Rubin, 2004; Beesley, Bartlett, Wolf, & Taylor, 2016; White & Royston, 2009). We, of course, hope that the imputation distribution is modeled correctly so that it does not introduce bias, since both the imputation model and the analysis model are based on Cox proportional hazards model which we assumed is the correct model for both cases.

As we will demonstrate in the next section, we can achieve a good imputation of the censored covariate values when we incorporate the observed outcome of interest Y, the temporal dependence of the censoring Δ affecting the true survival T (left-hand side of (2)), and the information provided by observed measures V of the covariate of interest X. At the end of the imputation process, for each observation i, the imputed value $\widetilde{X}_i$ is either equal to V_i if X_i was observed or to a function $g(V_i, Y_i, D_i, Z_i)$, if X_i was censored, where g is a known function, defined depending on the context as we show below.

Having a good idea of what such a substantive model will look like helps tremendously with the imputation process. The factorization (2) and the specification of the imputation distribution can be successfully completed in a way that it does not introduce bias in the parameter estimations of the covariate X. This ensures that inference based on the data available approximate the underlying likelihood of the observed data (Bartlett et al., 2015; Beesley et al., 2016; White & Royston, 2009).

To impute a randomly censored observation X_j, we replace it by an estimate of $E(X_j\, X_j > R_j, \boldsymbol{W})$, for $j = 1, \dots, N$, with $\mathbf{W} = (\mathbf{Z}, \boldsymbol{Z}^*, Y, \Delta)$ and $\boldsymbol{Z}^*$ is the vector of auxiliary covariates (if available) (Atem et al., 2017). Let $\boldsymbol{\tau} = max\{D_1X_1, \dots, D_NX_N\}$. We assumed that $\boldsymbol{\tau}$ belongs strictly on the support of X in the sense that, if X_i correspond to a censored observation, we considered it as an event X following Datta's assumption, that is $\boldsymbol{\tau}\ S(\boldsymbol{\tau}) = 0$ and $\int_{\tau}^{\infty} S(u)du = 0$ (Datta, 2005).

When the censoring in X is strictly independent of (Y, Δ) and in absence of covariates $\mathbf{Z}$ or if the censoring R is strictly independent from $\mathbf{Z}$ and T, the conditional expectation $E(X_j | X_j > R_j, \mathbf{Z})$ is just equal to $E(X_j | X_j > R_j)$, where

$$E\left(X_j | X_j > R_j\right) = S\left(R_j\right)^{-1} \int_{R_j}^{\tau} S(u)du + R_j, \tag{4}$$

where $S(h) = \int_h^{\infty} f(t)dt$ is the survival probability of X. We can estimate S using the Kaplan–Meier estimator $\hat{S}$, and linearly extrapolate $\hat{S}$, to approximate the values of S at censored observations, by using the mean between subsequent events (Atem et al., 2017; Datta, 2005).

Based on the trapezoidal approximation rule, the integral $\int_a^b S(u)du \approx \frac{1}{2}(b-a)[S(a) + S(b)]$, for $(a, b) \in \mathbf{R}^2$ and $a \neq b$. Hence, we can approximate $E(X_j | X_j > R_j)$ by

$$\sum_{i=1}^{n} I\left[V_{(i)} > R_j\right]\left[\frac{S\left(V_{(i)}\right) + S\left(V_{(i+1)}\right)}{2S\left(R_j\right)}\right]\left(V_{(i+1)} - V_{(i)}\right) + R_j,$$

where $V_{(1)} < V_{(2)} < \ldots < V_{(n)}$ are the ordered observed values of the variable $V = min(X, R)$.

When additional covariates $\mathbf{Z}$ and $\mathbf{Z}^*$ are available, we use a Cox proportional hazards model-based estimator $\hat{S}$ of S, that is, $\hat{S}(u) = \hat{S}_0(u)^{\exp\left(\alpha^\top \mathbf{W}_j\right)}$ for $u \in \mathbf{R}$, where $\hat{S}_0$ is the baseline survivor function. Specifically, we ran a Cox proportional hazards model for X given by $\lambda(x|\mathbf{w}) = \lambda_0(x) \exp(\alpha^\top \mathbf{w})$, where $\lambda(x|\mathbf{W})$ and $\lambda_0(x)$ are, respectively, the hazard and the baseline hazard for $X = x$ estimated at $\boldsymbol{W} = \mathbf{W}$ and $\boldsymbol{W} = 0$, respectively. Therefore,

$$E\left(X_j | X_j > R_j, \mathbf{W}\right) \approx \left(\left[2\hat{S}_0\left(R_j\right)\right]^{exp\left(\alpha^\top \mathbf{W}_j\right)}\right)^{-1} \times \sum_{i=1}^{n}\left\{I\left[V_{(i)} > R_j\right]\left[\hat{S}_0\left(V_{(i)}\right)^{exp\left(\alpha^\top \mathbf{W}_j\right)} + \hat{S}_0\left(V_{(i+1)}\right)^{exp\left(\alpha^\top \mathbf{W}_j\right)}\right]\left(V_{(i+1)} - V_{(i)}\right)\right\} + R_j. \quad (5)$$

We estimate the baseline survivor function $\hat{S}_0$ using the Breslow estimator (Breslow, 1972).

3.3 | Imputation algorithm

The variance estimation from the Cox proportional model, based on the above singly imputed data, does account for the uncertainty and the variability related to such an imputation process. It tends to underestimate the actual variance of the parameter of interest (Schafer, 1999; Little, 1992). To calibrate the estimated variance, we use the following algorithm:

1. Sample N units with replacement from the original data to form a bootstrap sample;

2. Use the bootstrap sample to replace each censored value X_j by the conditional mean $E(X_j|X_j > R_j, \boldsymbol{W}_j)$ as indicated in Section 3.2; then

   i. fit the model $h(t) = h_o(t)\exp\left(\beta_b X + \gamma_b^\top \mathbf{Z}\right)$

   ii. store the variance $\widehat{\mathrm{Var}}\left(\hat{\beta}_b\right)$

3. Repeat steps 1 to 2 B times ($B \geq 500$);

4. Calculate the estimated variance

$$\widehat{Var}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B}\left[\widehat{Var}\left(\hat{\beta}_b\right)\right],$$

where each $\widehat{\mathrm{Var}}\left(\hat{\beta}_b\right)$, for $b = 1, \ldots, B$ is the model-based variance from step 2.

The use of bootstrap resampling in our algorithm resemble the algorithm commonly used for the regression calibration method in measurement error theory (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Royston et al., 2006; Yi, 2017; Buonaccorsi, 2010; Tu & Greenwood, 2012).

3.4 | Multiple censored covariates

The foregoing method for random censored covariate can be extended to accommodate more than one randomly censored covariate, using the method for multivariate missing data analysis proposed by Buck (1960). As in the one covariate case, we are interested in assessing the relationship between the $\boldsymbol{K}$-vector of potentially censored covariates $\mathbf{X} = (X_1, \ldots, X_{\boldsymbol{K}})$ and the survival time T, adjusting for the effects of fully observed covariates $\boldsymbol{Z}$ through a Cox proportional hazards model

$$h(t) = h_o(t)\exp\left(\beta_1 X_1 + \cdots + \beta_K X_K + \gamma^\top \mathbf{Z}\right), \qquad (6)$$

where $h_o(t)$ is the baseline hazard and $(\beta_1, \ldots, \beta_K, \gamma^\top)$ are the unknown regression coefficients. We assume that model (6) is correctly specified and denote the vector of censoring indicators for $\mathbf{X}$ by $\mathbf{R} = (R_1, \ldots, R_{\boldsymbol{K}})$.

Following Buck (1960), we can factorize the distribution of $f(\mathbf{X}|T,\mathbf{Z})$ as $f(\mathbf{X},\mathbf{Z})\ f(T|\mathbf{X},\mathbf{Z})$

For censored value of the covariate X_K, we subsequently consider functions of the form

$$g_k\left(V_k, \left(X_1, D_1\right), \ldots, \left(X_{k-1}, D_{k-1}\right), \left(X_{k+1}, D_{k+1}\right), \ldots\left(X_K, D_K\right), (Y, \Delta), \mathbf{Z}\right) \qquad (7)$$

for some g_k, $k = 1, \ldots, \boldsymbol{K}$, and break down the imputation of censored covariates into a series of single-censored covariate procedures.

Similarly to the above one-covariate imputation, the conditional expectation $E(X_{Xj}|\ X_{Xj} > R_{Xj}, \boldsymbol{W}_{\{-k\}})$ is approximated by

$$\left(\left[2\hat{S}_0\left(R_{kj}\right)\right]^{exp\left(\alpha^\top \mathbf{W}_{\{-k\}j}\right)}\right)^{-1} \sum_{i=1}^{n} \left\{ I\left[V_{(ki)} > R_j\right] \times \left[\hat{S}_0\left(V_{(ki)}\right)^{exp\left(\alpha^T \mathbf{W}_{\{-k\}j}\right)} + \hat{S}_0\left(V_{(ki+1)}\right)^{exp\left(\alpha^\top \mathbf{W}_{\{-k\}j}\right)}\right] \times \left(V_{(ki+1)} - V_{(ki)}\right) \right\} + R_{kj}, \qquad (8)$$

where $V_{ki} = \min(X_{ki}, R_{ki})$, $k = 1, \ldots, \boldsymbol{K}$. The vector $\boldsymbol{W}_{\{-k\}}$ includes $\boldsymbol{Z}$, Y, Δ, $(X_1, \ldots, X_{k-1}, X_{k+1}, X_{\boldsymbol{K}})$ and $(D_1, \ldots, D_{k-1}, D_{k+1}, D_{\boldsymbol{K}})$, where $D_k = (D_{k1}, \ldots, D_{kN})$ and $D_{ki} = I(X_{ki} < R_{ki})$ represents the indicator of covariate censoring, for $k = 1, \ldots, \boldsymbol{K}$. Therefore, for each participant j, the imputed covariate value $\tilde{X}_{jk}$ will then be either equal to V_{jk} or equal to $E(X_{kj}|X_{kj} > R_{kj}, \boldsymbol{W}_{\{-k\}})$.

4 | SIMULATIONS

In this section, we demonstrate the characteristics of the proposed method via simulation studies. We generated different sets of data corresponding to different scenarios (and distributions) of right-censoring mechanism R for the censored variable X. In each of these different scenarios, we generated $M = 3{,}000$ simulated datasets of size N equal to 250, 500, and 750, respectively.

4.1 | Simulation setup

In each data set, we generated a fully observed covariate Z following the standard normal $Z \sim (0, 1)$ and the censored covariate $X \sim Weibull\left(\frac{3}{4}, \frac{1}{4}\right)$. In addition, we considered three different scenarios for the distribution of the censoring mechanism R of X:

1. no censoring, that is, the covariate X is fully observed;
2. $R \sim Weibull(1, 1)$ for 25% censoring;
3. $R \sim Weibull(1, 0.3)$ for 50% censoring.

The hazards of the survival events T were based on a proportional hazards model 1, where $h(t) = h_0(t)\exp(\beta X + \gamma Z)$ with $\beta = -1$ and $\gamma = 0.5$. Thus, the outcome was generated from $T \sim Weibull(1, 1/4\theta)$ where $\theta = \exp(-\beta X - \gamma Z)$. Finally, we generated the censoring variable $C \sim Weibull(1, q)$, where $q = 1$ and $q = 0.3$ to reach approximately 25% and 50% censored events.

Similarly, we considered the variable $Z \sim (0, 1)$ as above and generated also the censored covariates $X_1 \sim Weibull\left(\frac{3}{4}, \frac{1}{4}\right)$ and $X_2 \sim Weibull\left(\frac{1}{2}, \frac{1}{4}\right)$ under three different scenarios for the distribution of the censoring mechanisms R_1 and R_2 as follow:

1. no censoring, that is, X is fully observed;
2. $R_1 \sim Weibull(1, 1)$ for 25% censoring and $R_2 \sim Weibull\left(\frac{3}{4}, \frac{3}{4}\right)$ for 25% censoring;
3. $R_1 \sim Weibul(1, 0.3)$ for 50% censoring and $R_2 \sim Weibull\left(\frac{3}{4}, \frac{1}{4}\right)$ for 50% censoring.

In addition, we generated the outcome $T \sim Weibull(1, 1/4\theta)$, for $\theta = \exp(-\beta_1 X_1 - \beta_2 X_2 - \gamma Z)$ following the proportional hazards model 6 defined as $h(t) = h_0(t)\exp(\beta_{01} X_1 + \beta_{02} X_2 + \gamma Z)$ with $\beta_1 = -1$, $\beta_2 = 0.5$ and $\gamma = 0.5$. Its corresponding censoring random variable $C \sim Weibull(1, q)$ where $q = 1$ and $q = 0.3$ for, respectively, 25% and 50% censoring.

4.2 | Evaluation criteria

For each scenario of the censoring mechanisms R, R_1, or R_2 and for each of the $M = 3{,}000$ simulated data sets, we determined the estimates $\hat{\beta}_{km}$, $m = 1, \ldots, M$, of the parameters of interest β_k, $k = 1$ (resp. $k = 1, 2$), using the algorithm aforementioned with $B = 500$ bootstrap samples. We calculated the percentage $bias = 100\%\left(\bar{\hat{\beta}}_k - \beta_k\right)/\beta_k$, where $\bar{\hat{\beta}}_k = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}_{km}$; the Monte Carlo simulation standard error, that is, the empirical standard error (ESE) of the estimate of interest over all M simulated data sets; the model-based standard error (SE) which is the average of all standard errors from each fitted model; the mean squared error (MSE); and the 95% coverage probability, that is, the proportion of the time that the 95 % confidence interval contains the true parameter value β_k, $k = 1$ (resp. $k = 1, 2$).

We compared the results from our proposed method—the conditional imputation with bootstrap (CI-B)— to those from (1) the complete-case analysis (CC) to highlight the impact of censoring on the censored covariate and (2) the conditional imputation without bootstrap (CI), to show the importance of bootstrap variance estimation in accounting for the uncertainty related to the imputation scheme.

The most consistent and efficient approach will be the one that has the smallest bias, a model-based standard error similar to the simulation error, the smallest MSE, and a coverage probability close to 0.95. Naturally, we expected the scenario with no censoring (full data) to yield results with the most consistent and efficient model parameters. Therefore, the efficiency of our proposed method was then assessed based on how close it is from the non-censored covariate scenario compared to the other two methods.

4.3 | Simulation results

Tables 1–3 report the results of our simulation studies. An inspection of these tables indicate that the performance of the different methods depends on both the sample size and the percentage of censored covariate(s). As expected, for each method, given a percentage of censored covariate, the bias decreases as the sample size increases. Also, for a given sample size, the bias increases as the percentage of censored covariate increases from 25% to 50%. All three methods result in reasonable coverage probabilities.

For a sample size $N = 250$, all three methods were slightly biased and comparable. However, the standard error of the complete-case analyses were high and increased tremendously as the censoring rate increased. This explains the high MSEs for the complete-case analysis compared to the other two imputation methods. The conditional imputation with bootstrap performed better than the complete-case analysis since the relative efficiency, measured by MSE, was higher in the complete-case than both imputation approaches. The efficiency for conditional imputation with bootstrap was more pronounced as compared to the complete-case analysis when multiple variables were subject to random censoring (Table 3).

Most of the 95% coverage probabilities were adequate. However, for smaller sample size $N = 250$, with heavy censoring, that is for 50% censoring, the coverage probabilities fell below 95 % for all the methods under both Kaplan–Meier and Cox proportional hazards model estimation. In fact, the coverage probabilities even dropped below 94% for both the complete-case and conditional imputation methods for 50% censoring under Kaplan–Meier approximation. As the sample size increased to $N = 500$ and $N = 750$, the bias decreased and the coverage probabilities became close to (and sometimes higher than) the nominal 95% level for all three methods.

Overall, since deleting censored observation reduces precision, the simulation standard errors and the model-based standard errors for the complete-case analyses were the largest compared to the other two imputation methods especially when two variables are subject to censoring. The simulation results show that the complete-case method yielded the highest bias, standard error and MSE, especially under heavy censoring of multiple variables. This is the result of reduction in precision since censored observations are deleted. The higher the percentage of censoring, the higher the standard error in the complete-case analysis. The increase in standard error for the complete-case (as compared to the other two methods) was even more pronounced when censoring increased to 50% and with two variables subject to censoring.

The conditional imputation based on the Cox model performs better than the conditional imputation based on Kaplan–Meier for $n = 750$. For larger sample size, the study is well-powered to estimate additional parameters that can be included in the model. Furthermore, in the context of missing data, (Meng, 1994) recommend that all the necessary variables—including auxilliary variables and the outcome—should be added to the imputation model. This is extremely crucial when data are missing at random. Similarly, when dealing with censored data, it is important to include all variables in the imputation model (including the outcome and the corresponding censoring indicator). The imputation based on Kaplan–Meier does not perform as well as that the imputation based on the Cox model because the latter includes all the variables available in the analysis phase and thus improves the imputation.

The simulation error of both conditional imputations—either based on Kaplan–Meier or Cox proportional hazards model—were always greater than the model-based error because both imputation methods did not adjust for the extra uncertainty arising from the imputation process. Moreover, the CI method results in the smallest standard error since in estimating its corresponding standard error the uncertainty related to the imputation process is not taken into consideration. However, this is improved through the bootstrap resampling in our proposed CIB method. This adjustment in standard error resulted in larger MSE in the both conditional imputation with bootstrap under Kaplan–Meier and Cox proportional hazards as compared to conditional imputation for one or two censored covariates imputation.

5 | APPLICATION

To illustrate the proposed method, we considered the data from the Framingham Heart Study (FHS) introduced in Section 2. We assessed the association between age at cardiovascular event in offspring and age of onset of cardiovascular event in parents, after controlling for gender, body mass index (BMI ≤ 30 or BMI > 30), and wine intake (yes or no).

One major assumption, for illustrative purposes, is that we have treated death as an independent censoring event for onset of CVD in our data analysis. If death is instead a competing risk then the independence assumption would be violated. While we cannot test for independent censoring, we might be concerned that death prior to CVD would not always qualify as an independent censoring event. In this case, we need to view our Cox models as a cause model for the cause-specific hazard for CVD that are conditional on being alive. In applications in which this is a major concern, competing risk methodology could be used to estimate a cumulative incidence function and employ a survival model that adjusts for the competing risks.

We performed two sets of analyses. First, we ran two separate Cox proportional hazards models to assess the association between age at cardiovascular events in offspring and maternal (resp. paternal) age at onset of cardiovascular event. Second, we investigated the joint association of both paternal and maternal age at onset of cardiovascular events on offspring age at first cardiovascular event in a single Cox proportional hazards model.

In the data set restricted to mothers and offspring, a total of 907 out of 1,401 mothers and 388 out of the 1,401 offspring participants experienced a cardiovascular event. The percentages of censored measures in each these groups were 35.26% and 72.30%. On the other hand, of the 1,221 fathers in the father-offspring data set, 909 of them experienced a cardiovascular event (censoring percentage equal to 25.55%). In addition, there were 283 out 1,221 offspring participants who had a cardiovascular event (thus 76.82% censored observations).

Table 4 summarizes the results of the two separate analyses. All the methods led to significant association between cardiovascular event in offspring and parents. This indicates that offspring participants whose parents had a early onset of cardiovascular disease, have higher hazards of a cardiovascular event. As expected, the complete-case analysis was the least powerful method since deleting censored observations reduced the sample size.

We also assessed the association between parental and offspring age at onset of a cardiovascular event by running a Cox proportional hazards model that includes the two censored covariates. The final data set contained a total of 1,141 observations that had both parents' age at onset of CVD. Of the 1,141 parents, only 555 mothers and fathers experience CVD. Thus, for the complete-case analysis, 586 (i.e., 51.36%) observations were discarded because of censoring.

Table 5 shows that there is a tremendous loss in power using the complete-case analysis as a consequence of such a considerable deletion of valuable data. Moreover, the results from Table 5 based on the complete-case analysis show a non-significant association between father age of onset of CVD and offspring age of onset of CVD (as opposed to the results obtained in Table 4 where we ran a separate model for paternal's age of onset of CVD event).

However, both the CI and CI-B results in significant association between both parents age of onset of CVD and offspring age of onset of CVD; these results are compatible to those obtained from Table 4 when running two separate models.

6 | DISCUSSION

In this paper, we have proposed an imputation method for randomly censored covariates that uses either a Kaplan–Meier estimation or a Cox proportional hazards model (when there are additional baseline or auxiliary covariates) to input censored observations. The imputed data is then use to assess the effects of these convariates on a time-to-event outcome through an additional Cox regression proportional hazards model.

Our imputation method is similar to the conditional imputation of Atem et al. (2017) for linear regression model and in the spirit of the conditional imputation of Little for missing data (Little, 1992). It consists in replacing censored covariate observations by their corresponding conditional mean values, given all the other covariates as well as the time-to-event outcome and auxiliary variables (when available). Each conditional mean is determined based on either the Kaplan–Meier approach or a Cox proportional hazards model of the censored covariate(s) using a trapezoidal integral approximation rule (Atkinson, 2008). Furthermore, the way we have leverage the time-to-event outcome in our imputation process is different and more delicate than Atem et al. (2017) did for linear regression model with a continuous outcome, due to the nature of the time-to-event outcome since it is subject to random censoring as well.

Unlike the complete-case analysis, which fails to capitalize on the available information in the data on the censored observations, our proposed imputation method allows us to not only replace censored values with the best approximation possible, but also to retain all the available information in a more complete data set. The conditional imputation without bootstrap method also provides unbiased parameter estimates, but underestimates the standard error as it does not make up for reduced variability in error due to imputation scheme. However, this is improved through the bootstrap resampling in our proposed method. Our method yielded valid inference in terms of the parameter estimates, the bias, and the mean-squared error. Finally, by analyzing the Framingham Heart Study, we showed that when censored observations are properly accounted for using our imputation method, we can have a good estimate of the effect of the age of onset of a parental cardiovascular event on the age of cardiovascular disease in their offspring.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their insightful comments, inspiring suggestions, and meaningful discussions, which have helped improve the quality of this paper tremendously. The work by Roland A. Matsouaka was supported, in part, by the grant UL1TR001117 awarded by the National Institutes of Health. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official view of the National Institutes of Health.

Funding information

National Institutes of Health, Grant/Award Number: UL1TR001117

REFERENCES

Atem F, & Matsouaka RA (2017). Linear regression model with a randomly censored predictor: Estimation procedures. arXiv:1710.08349.

Atem FD, Qian J, Maye JE, Johnson KA, & Betensky RA (2016). Multiple imputation of a randomly censored covariate improves logistic regression analysis. Journal of Applied Statistics, 43, 1–112886–2896.

Atem FD, Qian J, Maye JE, Johnson KA, & Betensky RA (2017). Linear regression with a randomly censored covariate: Application to an alzheimer's study. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(2), 313–328.

Atem FD, Sampene E, & Greene TJ (2017). Improved conditional imputation for linear regression with a randomly censored predictor. Statistical Methods in Medical Research, 28, 432–444. [PubMed: 28830304]

Atkinson KE (2008). An introduction to numerical analysis. New York: John Wiley & Sons.

Austin PC, & Hoch JS (2004). Estimating linear regression models in the presence of a censored independent variable. Statistics in Medicine, 23(3), 411–429. [PubMed: 14748036]

Bartlett JW, Seaman SR, White IR, Carpenter JR, & Initiative ADN. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research, 24(4), 462–487. [PubMed: 24525487]

Beesley LJ, Bartlett JW, Wolf GT, & Taylor JM (2016). Multiple imputation of missing covariates for the cox proportional hazards cure model. Statistics in Medicine, 35, 4701–4717. [PubMed: 27439726]

Bernhardt PW, Wang HJ, & Zhang D (2014). Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. Computational Statistics & Data Analysis, 69, 81–91.

Breslow NE (1972). Contribution to the discussion of the paper by Dr Cox. Journal of the Royal Statistical Society, Series B, 34(2), 216–217.

Buck SF (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society. Series B (Methodological), 22, 302–306.

Buonaccorsi JP (2010). Measurement error: Models, methods, and applications. Boca Raton, FL: CRC Press.

Carroll RJ, Ruppert D, Stefanski LA, & Crainiceanu CM (2006). Measurement error in nonlinear models: A modern perspective, Boca Raton, FL: CRC press.

Chen G, & Levy D (2016). Contributions of the Framingham heart study to the epidemiology of coronary heart disease. JAMA Cardiology, 1(7), 825–830. [PubMed: 27464224]

Chen Q, Wu H, Ware LB, & Koyama T (2014). A Bayesian approach for the cox proportional hazards model with covariates subject to detection limit. International Journal of Statistics in Medical Research, 3(1), 32. [PubMed: 24772198]

Cole SR, Chu H, Nie L, & Schisterman EF (2009). Estimating the odds ratio when exposure has a limit of detection. International Journal of Epidemiology, 38(6), 1674–1680. [PubMed: 19667054]

Collins LM, Schafer JL, & Kam C-M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330. [PubMed: 11778676]

D’Angelo G, & Weissfeld L (2008). An index approach for the cox model with left censored covariates. Statistics in Medicine, 27(22), 4502–4514. [PubMed: 18407573]

Datta S (2005). Estimating the mean life time using right censored data. Statistical Methodology, 2(1), 65–69.

Dinse GE, Jusko TA, Ho LA, Annam K, Graubard BI, Hertz-Picciotto I, Miller FW, Gillespie BW, & Weinberg CR (2014). Accommodating measurements below a limit of detection: A novel application of cox regression, American Journal of Epidemiology, 179(8), 1018–1024. [PubMed: 24671072]

Fitzsimons GJ (2008). Death to dichotomizing. Journal of Consumer Research, 35(1), 5–8.

Helsel DR (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. Chemosphere, 65(11), 2434–2439. [PubMed: 16737727]

Helsel DR (2011). Statistics for censored environmental data using Minitab and R, Vol. 77 New York: John Wiley & Sons.

Helsel DR (2005). Nondetects and data analysis Statistics for censored environmental data. New York: Wiley-Interscience.

Kalbfleisch JD, & Prentice RL (2011). The statistical analysis of failure time data, Vol. 360 New York: John Wiley & Sons.

Klein JP, & Moeschberger ML (2005). Survival analysis: Techniques for censored and truncated data. New York: Springer Science & Business Media.

Kong S, & Nan B (2016). Semiparametric approach to regression with a covariate subject to a detection limit. Biometrika, 103(1), 161–174.

Langohr K, Gómez G, & Muga R (2004). A parametric survival model with an interval-censored covariate. Statistics in Medicine, 23(20), 3159–3175. [PubMed: 15449329]

Lee JS, Cole SR, Richardson DB, Dittmer DP, Miller WC, Moore RD, … Eron JJ Jr; Center for AIDS Research Network of Integrated Clinical Systems (2017). Incomplete viral suppression and mortality in hiv patients after antiretroviral therapy initiation. Aids, 31(14), 1989–1997. [PubMed: 28650383]

Lee S, Park S, & Park J (2003). The proportional hazards regression with a censored covariate. Statistics & Probability Letters, 61(3), 309–319.

Lipsitz S, Parzen M, Natarajan S, Ibrahim J, & Fitzmaurice G (2004). Generalized linear models with a coarsened covariate. Journal of the Royal Statistical Society: Series C (Applied Statistics), 53(2), 279–292.

Little RJ (1992). Regression with missing x's: A review. Journal of the American Statistical Association, 87(420), 1227–1237.

Meng X-L (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 9, 538–558.

Rathouz PJ (2007). Identifiability assumptions for missing covariate data in failure time regression models. Biostatistics, 8(2), 345–356. [PubMed: 16840561]

Rigobon R, & Stoker TM (2009). Bias from censored regressors. Journal of Business & Economic Statistics, 27(3), 340–353.

Royston P, Altman DG, & Sauerbrei W (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. Statistics in Medicine, 25(1), 127–141. [PubMed: 16217841]

Rubin DB (2004). Multiple imputation for nonresponse in surveys, Vol. 81 New York: John Wiley & Sons.

Sattar A, Sinha SK, & Morris NJ (2012). A parametric survival model when a covariate is subject to left-censoring. Journal of Biometrics & Biostatistics, S3.

Sattar A, Sinha SK, Wang X-F, & Li Y (2015). Frailty models for pneumonia to death with a left-censored covariate. Statistics in Medicine, 34(14), 2266–2280. [PubMed: 25728821]

Schafer JL (1999). Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1), 3–15. [PubMed: 10347857]

Tsimikas JV, Bantis LE, & Georgiou SD (2012). Inference in generalized linear regression models with a censored covariate. Computational Statistics & Data Analysis, 56(6), 1854–1868.

Tu Y-K, & Greenwood DC (2012), Modern methods for epidemiology. New York: Springer Science & Business Media.

Wang HJ, & Feng X (2012). Multiple imputation for m-regression with censored covariates. Journal of the American Statistical Association, 107(497), 194–204.

White IR, & Royston P (2009). Imputing missing covariate values for the cox model. Statistics in Medicine, 28(15), 1982–1998. [PubMed: 19452569]

Wu H, Chen Q, Ware LB, & Koyama T (2012). A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: An application to acute lung injury. Journal of Applied Statistics, 39(8), 1733–1747. [PubMed: 23049157]

Yi GY (2017). Statistical analysis with measurement error or misclassification: Strategy, method and application, New York: Springer Science & Business Media.

Yue YR, & Wang X-F (2016). Bayesian inference for generalized linear mixed models with predictors subject to detection limits: An approach that leverages information from auxiliary variables. Statistics in Medicine, 35(10), 1689–1705. [PubMed: 26643287]

Zhou H, & Pepe MS (1995). Auxiliary covariate data in failure time regression. Biometrika, 82(1), 139–149.

TABLE 1

Imputation of a censored covariate based on Kaplan–Meier Estimation

		Light censoring (25% censoring)					**Heavy censoring (50% censoring)**				
Size	**Method**	**Bias***	**SE**	**ESE**	**MSE**	**CP**	**Bias***	**SE**	**ESE**	**MSE**	**CP**
250	Full	−0.80	0.245	0.250	0.060	96.2	−0.80	0.245	0.250	0.060	96.2
	CC	0.81	0.691	0.720	0.478	95.4	−4.10	0.903	0.903	0.817	92.2
	CI	2.91	0.362	0.420	0.132	95.8	3.77	0.464	0.700	0.216	91.3
	CI-B	2.91	0.405	0.440	0.165	96.1	3.77	0.620	0.642	0.386	92.0
500	Full	0.26	0.170	0.175	0.029	96.9	0.26	0.170	0.175	0.029	96.9
	CC	0.71	0.497	0.498	0.247	96.4	4.00	0.642	0.653	0.414	95.0
	CI	2.61	0.226	0.297	0.052	96.0	−3.48	0.310	0.350	0.097	95.2
	CI-B	2.61	0.276	0.278	0.077	96.1	−3.48	0.338	0.349	0.116	95.4
750	Full	−0.16	0.138	0.138	0.019	97.4	−0.16	0.138	0.138	0.019	97.4
	CC	0.35	0.409	0.411	0.167	97.0	−2.93	0.521	0.524	0.273	97.0
	CI	0.84	0.195	0.224	0.038	96.2	3.52	0.249	0.328	0.063	96.2
	CI-B	0.84	0.229	0.233	0.053	96.2	3.52	0.276	0.278	0.079	96.3

Note. Full = full data; CC = complete-case; CI = conditional imputation; CI-B = CI with bootstrap; SE = (model-based) standard error; ESE = empirical standard error; CP = estimated coverage probability of 95% confidence intervals.

* Relative bias percentage $= \left(\frac{\hat{\beta} - \beta}{\beta}\right) \times 100\%$.

TABLE 2

Imputation of a censored covariate based on Cox proportional hazards model

		Light censoring (25% censoring)					Heavy censoring (50% censoring)				
Size	**Method**	**Bias***	**SE**	**ESE**	**MSE**	**CP**	**Bias***	**SE**	**ESE**	**MSE**	**CP**
250	Full	1.61	0.253	0.254	0.064	96.2	1.61	0.253	0.254	0.064	96.2
	CC	−2.84	0.379	0.389	0.144	96.1	4.41	0.782	0.800	0.613	93.2
	CI	2.84	0.263	0.331	0.070	96.0	4.11	0.449	0.501	0.203	92.3
	CI-B	2.84	0.323	0.365	0.105	96.1	4.11	0.488	0.540	0.240	94.0
500	Full	−0.74	0.171	0.183	0.029	97.1	−0.74	0.171	0.183	0.029	97.1
	CC	0.96	0.266	0.279	0.071	96.1	4.19	0.642	0.669	0.414	95.1
	CI	0.91	0.183	0.270	0.033	96.1	4.10	0.313	0.341	0.100	94.0
	CI-B	0.91	0.221	0.230	0.049	96.2	4.10	0.329	0.341	0.110	95.1
750	Full	−0.07	0.138	0.141	0.019	97.6	−0.07	0.138	0.141	0.019	97.6
	CC	0.74	0.215	0.226	0.046	97.0	−2.24	0.525	0.542	0.276	96.0
	CI	0.72	0.148	0.194	0.022	96.2	2.05	0.240	0.284	0.058	95.0
	CI-B	0.72	0.175	0.180	0.031	96.2	2.05	0.288	0.331	0.083	95.6

Note. Full = full data; CC = complete-case; CI = conditional imputation; CI-B = CI with bootstrap; SE = (model-based) standard error; ESE = empirical standard error; CP = estimated coverage probability of 95% confidence intervals.

* Relative bias percentage $= \left(\frac{\hat{\beta} - \beta}{\beta}\right) \times 100\%.$

TABLE 3

Imputation of two censored covariates based on Cox proportional hazards model

		X_1					X_2				
Size	**Method**	**Bias***	**SE**	**ESE**	**MSE**	**CP**	**Bias***	**SE**	**ESE**	**MSE**	**CP**
250		(1a) Light censoring (25% censoring)									
	Full	0.808	0.254	0.254	0.065	0.963	0.004	0.129	0.129	0.017	0.963
	CC	0.014	0.477	0.504	0.228	0.957	0.004	0.412	0.450	0.170	0.949
	CI	0.009	0.301	0.502	0.091	0.950	0.004	0.174	0.220	0.030	0.960
	CI-B	0.009	0.2911	0.311	0.085	0.960	0.004	0.178	0.191	0.032	0.949
		(1b) Heavy censoring (50% censoring)									
	Full	0.008	0.254	0.254	0.065	0.963	0.004	0.129	0.129	0.017	0.963
	CC	0.039	1.070	1.205	1.146	0.925	0.094	1.103	1.224	1.225	0.921
	CI	0.010	0.380	0.432	0.145	0.950	0.007	0.271	0.413	0.073	0.960
	CI-B	0.010	0.399	0.415	0.159	0.957	0.007	0.271	0.300	0.073	0.949
500		(2a) Light censoring (25% censoring)									
	Full	0.004	0.179	0.182	0.032	0.964	0.003	0.091	0.092	0.008	0.968
	CC	0.013	0.330	0.336	0.109	0.964	0.004	0.268	0.261	0.072	0.967
	CI	0.004	0.200	0.353	0.040	0.954	0.003	0.116	0.171	0.013	0.967
	CI-B	0.004	0.216	0.222	0.047	0.964	0.003	0.137	0.142	0.019	0.950
		(2b) Heavy censoring (50% censoring)									
	Full	0.004	0.179	0.182	0.032	0.964	0.003	0.091	0.092	0.008	0.968
	CC	0.028	0.722	0.755	0.522	0.944	0.009	0.706	0.749	0.499	0.954
	CI	0.006	0.256	0.306	0.066	0.950	0.005	0.168	0.288	0.028	0.964
	CI-B	0.006	0.268	0.295	0.072	0.951	0.005	0.191	0.222	0.037	0.950
750		(3a) Light censoring (25% censoring)									
	Full	0.004	0.144	0.149	0.021	0.965	0.002	0.072	0.071	0.005	0.968
	CC	0.013	0.268	0.285	0.072	0.964	0.003	0.216	0.224	0.047	0.967
	CI	0.004	0.160	0.183	0.026	0.960	0.003	0.092	0.209	0.008	0.955
	CI-B	0.004	0.168	0.183	0.028	0.964	0.003	0.094	0.121	0.008	0.955
		(3b) Heavy censoring (50% censoring)									
	Full	0.004	0.144	0.149	0.021	0.965	0.002	0.072	0.071	0.005	0.968
	CC	0.009	0.580	0.590	0.336	0.960	0.007	0.551	0.594	0.304	0.954
	CI	0.005	0.204	0.245	0.042	0.960	0.007	0.130	0.244	0.017	0.964
	CI-B	0.005	0.210	0.230	0.044	0.957	0.007	0.143	0.175	0.020	0.951

Note. Full = full data; CC = complete-case; CI = conditional imputation; CI-B = CI with bootstrap; SE = (model-based) standard error; ESE = empirical standard error; CP = estimated coverage probability of 95% confidence intervals.

* Relative bias $\% = \left(\frac{\hat{\beta} - \beta}{\beta}\right) \times 100\%.$

TABLE 4

Association between maternal (resp. paternal) age and offspring age of onsets of cardiovascular event

Method	Estimate	SE	*p*-Value
Maternal age			
CC	−0.04	7×10^{-3}	$< 10^{-4}$
CI	−0.03	6×10^{-3}	$< 10^{-4}$
CI-B	−0.03	6.1×10^{-3}	$< 10^{-4}$
Paternal age			
CC	−0.02	8.3×10^{-3}	7×10^{-3}
CI	−0.02	7×10^{-3}	10^{-3}
CI-B	−0.02	7.2×10^{-3}	10^{-3}

Note. CC = complete-case; CI = conditional imputation; CI-B = conditional imputation with bootstrap.

TABLE 5

Association between parental age and offspring age of onsets of cardiovascular event; including both parents' ages at first onset of a cardiovascular event in a single model

Method	Mother	SE	p-value	Father	SE	p-value
CC	−0.05	1.1×10^{-2}	$< 10^{-4}$	2×10^{-3}	1.1×10^{-2}	0.85
CI	−0.03	6.5×10^{-3}	$< 10^{-4}$	−0.02	6.4×10^{-3}	4×10^{-3}
CI-B	−0.03	6.6×10^{-3}	$< 10^{-4}$	−0.02	6.7×10^{-3}	0.02

CC = complete-case; CI = conditional imputation; CI-B = CI with bootstrap.