Annals of the Institute of Statistical Mathematics (2020) 72:969–996
https://doi.org/10.1007/s10463-019-00716-4

# Large sample results for frequentist multiple imputation for Cox regression with missing covariate data

**Frank Eriksson[1] · Torben Martinussen[1] · Søren Feodor Nielsen[2]**

Received: 31 July 2018 / Revised: 7 March 2019 / Published online: 4 April 2019


**Abstract**
Incomplete information on explanatory variables is commonly encountered in studies of possibly censored event times. A popular approach to deal with partially observed covariates is multiple imputation, where a number of completed data sets, that can be analyzed by standard complete data methods, are obtained by imputing missing values from an appropriate distribution. We show how the combination of multiple imputations from a compatible model with suitably estimated parameters and the usual Cox regression estimators leads to consistent and asymptotically Gaussian estimators of both the finite-dimensional regression parameter and the infinite-dimensional cumulative baseline hazard parameter. We also derive a consistent estimator of the covariance operator. Simulation studies and an application to a study on survival after treatment for liver cirrhosis show that the estimators perform well with moderate sample sizes and indicate that iterating the multiple-imputation estimator increases the precision.

**Keywords** Asymptotic distribution · Coarsened data · Semiparametric · Survival · Variance estimator

## 1 Introduction

The possible effect of prognostic factors $X$ on a censored time-to-event outcome is often modeled using the Cox model (Cox 1972), specified by the conditional hazard function

$$\alpha(t|X = x) = \alpha(t) \exp(\beta^{\top} x), \tag{1}$$

✉ Frank Eriksson
eriksson@sund.ku.dk

1 Section of Biostatistics, Department of Public Health, University of Copenhagen, 1014 Copenhagen, Denmark

2 Center for Statistics, Department of Finance, Copenhagen Business School, Solbjerg Plads 3, 2000 Frederiksberg, Denmark

where $\beta$ denotes the regression parameter and $\alpha(t)$ is the baseline hazard function, which is not further specified. Statistical inference in the Cox model with no missing data is well established, but in practice some values of $X$ may be missing. Literature studying possible solutions to this problem is extensive. One solution is to use inverse probability weighted estimators (Pugh et al. 1993), but these may suffer from low efficiency. Augmenting may improve the efficiency but the optimal augmenting function may be difficult to estimate in practice. Another way of improving the efficiency of inverse probability weighted estimators is to estimate the weights nonparametrically as shown by Qi et al. (2005). However, when having several covariates nonparametric estimation is affected by the curse of dimensionality and the higher-order kernels used by Qi et al. (2005) may result in estimated selection probabilities outside the unit interval. An alternative direction is to use full likelihood based methods (Chen and Little 1999; Martinussen 1999; see also Chen 2002; Herring and Ibrahim 2001). This leads to efficient estimators but requires specialized programming. Another class of methods, that are popular in practice, are imputation methods, where missing data are replaced by suitably generated “best guesses”, which can then be analyzed by standard software. Multiple imputation methods, where the imputation and estimation process is repeated a number of times and the estimators subsequently combined, are particularly popular, in part because the simulation noise may be diminished by repeated imputation. It has been stressed in the literature that the imputations should be done with care and that the response must be included in the imputation model, see Sterne et al. (2009). This has created some confusion when dealing with survival data where the response is censored. The problem was investigated in some detail by White and Royston (2009). Taking their approach, however, may lead to models that are incompatible which in turn may result in inconsistent estimates as shown by Bartlett et al. (2015). Bartlett et al. (2015) also show how rejection sampling may be used to generate the imputations that ensure model compatibility. In their paper, they devised a Bayesian multiple imputation procedure that seems to work well judging from their numerical results. Unfortunately, they did not establish large sample results for this procedure.

Although multiple imputation is widely used in practice for analyzing survival data with the Cox model, there exists, to the best of our knowledge, no formal results justifying its appropriateness. General asymptotic results for multiple imputation estimators in parametric models, such as those established by Wang and Robins (1998) and Robins and Wang (2000)(see also Tsiatis 2006), rely on stochastic equicontinuity of a process, which is not stochastic equicontinuous in this setting. Thus, the large sample properties of the estimators are unclear, and the validity of the suggested standard error estimators is unknown. This is unfortunate as it may invalidate scientific conclusions based on such analysis.

In this paper, we study the properties of multiple imputation estimators based on imputations from a compatible model. Such imputations may be generated using rejection sampling as suggested by Bartlett et al. (2015). We focus on what Tsiatis (2006) call frequentist multiple imputation, i.e., the case where the imputation model is based on a consistent and asymptotically linear initial estimator. Estimators of the finite-dimensional regression parameter and the infinite-dimensional cumulative baseline hazard parameter are shown to be $\sqrt{n}$-consistent and weak convergence is established.

Furthermore, we provide a consistent estimator of the asymptotic variance for the estimator of the regression parameter as well as a consistent estimator for the covariance operator for the estimator of the cumulated baseline hazard. Hence, our results provide the necessary justification for drawing correct statistical inference when using multiple imputation in Cox regression. Finally, we discuss how to improve on the multiple imputation estimators using a simple iterative scheme. The finite sample performance of the proposed estimators is investigated using simulations, and we further apply them in a study on survival after treatment for liver cirrhosis.

## 2 Frequentist multiple imputation for Cox regression

Let $X$ denote a $p$-dimensional vector of prognostic covariates that are partially missing for some individuals. Assume that the distribution of the event time $\tilde{T}$ given $X$ is governed by the Cox model (1). $\tilde{T}$ may be censored by $U$ and we only observe the minimum of the two $T = \tilde{T} \wedge U$ and the event indicator $\Delta = I(\tilde{T} \leq U)$. We assume that $\tilde{T}$ and $U$ are independent given the always observed part of $X$. Assume that $T$ is observed on the finite time interval $[0, \tau]$. The full data, denoted by $Z_1, \ldots, Z_n$, are independent realizations of $Z = (T, \Delta, X)$ with density

$$\alpha(t)^{\delta} \exp(\delta\beta^{\top}x) \exp\left\{-A(t)\exp(\beta^{\top}x)\right\} \alpha_U(t|x)^{1-\delta} pr(U > t|x) p_X(x, \theta),$$

for $z = (t, \delta, x)$, where the density of $X$, $p_X(x, \theta)$, is known up to the $q$-dimensional parameter $\theta$, and $A(t) = \int_0^t \alpha(s)\, ds$ is the integrated baseline hazard function. Let $\phi = (\beta, A, \theta)$ and let $\phi_0$ denote the true parameter. Assume that the censoring hazard $\alpha_U(t|X)$ does not depend on $\phi$ or partially unobserved covariates. Under full data, we would estimate $\beta_0$ by Cox's partial likelihood estimator and $A_0(t)$ by the corresponding Breslow estimator.

The data is assumed missing (or coarsened) at random, and the observed data is $\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$, where $\mathcal{C}$ denotes the missingness pattern and $G_{\mathcal{C}}(Z) = \{T, \Delta, G_{X,\mathcal{C}}(X)\}$ with $G_{X,r}(x)$ denoting the observed part of $x$ under missingness pattern $\mathcal{C} = r$, using a similar notation as in Tsiatis (2006). Thus with missing data we may let $\mathcal{C} = r$ be a vector of response indicators, i.e., a vector of zeros and ones denoting (by 1) which components of $X$ are observed and which are missing (corresponding to 0), or as in Tsiatis (2006) simply a number indicating which missingness pattern we observe for this observation. $G_{X,\mathcal{C}}(X)$ may then be just the actually observed values. Our notation and results also apply to data that are coarsened at random; see Jacobsen and Keiding (1995) for examples of how to represent coarsened data by $\{\mathcal{C}, G_{X,\mathcal{C}}(X)\}$. In the appendix we argue that the part of the density function adhering to censoring can be ignored when estimating $\phi_0$. Furthermore, both the censoring mechanism and the missing data mechanism may be ignored when imputing the missing covariates.

In this paper, we consider what Tsiatis (2006) refers to as frequentist multiple imputation and Wang and Robins (1998) call "type B". For each observed data $\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, we wish to sample at random from the conditional distribution of $X$ given the observed data with density $p_{X|\mathcal{C},G}\{x|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \phi_0\}$, but to do so we need to estimate the parameter $\phi_0$. We assume that an initial consistent and asymptotically linear estimator $\hat{\phi}^I$ is available. We then sample at random from $p_{X|\mathcal{C},G}\{x|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i), \hat{\phi}^I\}$ $m$ times to obtain random quantities $X_{ij}(\hat{\phi}^I)$, $j = 1, \ldots, m$, $i = 1, \ldots, n$. One way of sampling from this distribution is to use rejection sampling, i.e., by generating proposals from another distribution and accepting these with a suitable probability to make the resulting sample a sample from the desired distribution. How to do this when the substantive model is a Cox model has been described by Bartlett et al. (2015), who used conditional distributions derived from the distribution of $X$ to generate proposals. However, our results do not rely on how the imputations are generated as long as the imputations have the correct conditional distribution.

Standard Cox regression analysis on the $j$th set of imputed full data yields the estimators $\{\hat{\beta}_j, \hat{A}_j(t)\}$. The multiple-imputation estimators are

$$\hat{\beta} = m^{-1} \sum_{j=1}^{m} \hat{\beta}_j, \quad \hat{A}(t) = m^{-1} \sum_{j=1}^{m} \hat{A}_j(t), \tag{2}$$

where $\hat{\beta}_j$ is the maximizer of Cox's partial likelihood function based on the $j$th set of imputations and $\hat{A}_j(t)$ the corresponding Breslow estimator.

## 3 Asymptotics

In order to present our result regarding the asymptotic distribution of $\hat{\beta}$ and $\{\hat{A}(t)\}_{t\in[0;\tau]}$, we need to introduce some notation. The asymptotic representation of the full-data efficient score for $\beta$ evaluated at $\phi_0$ is

$$S_{\text{eff}}^F(Z) = \int_0^\tau \left\{X - \frac{s_1(t)}{s_0(t)}\right\} \mathrm{d}M^F(t, Z),$$

where $\mathrm{d}M^F(t, Z) = \mathrm{d}N(t) - Y(t)\exp(\beta_0^\top X)\alpha_0(t)\mathrm{d}t$ with $N(t) = I(T \leq t, \Delta = 1)$ and $Y(t) = I(T > t)$, $s_k(t) = E[S_k\{t, Z(\phi_0), \beta_0\}]$, and $S_k(t, Z, \beta) = Y(t)X^{\otimes k}\exp(\beta^\top X)$, $k = 0, 1, 2$.

Let the continuous linear operator $\mathcal{S}_\phi(z) : \mathbb{R}^p \times \ell^\infty[0, \tau] \times \mathbb{R}^q \mapsto \mathbb{R}$ denote the Hadamard derivative of $\log\{\tilde{p}_Z(z, \phi)\}$, where $\tilde{p}_Z(z, \phi) = \exp\{\delta\beta^\top x - A(t)\exp(\beta^\top x)\}p_X(x, \theta)$, at $\phi$ (van der Vaart 1998, Section 20.2). The derivative at $\phi_0$ in the direction $(\hat{\phi}^I - \phi_0)$ is given by

$$\begin{aligned}\mathcal{S}_{\phi_0}(z)\left(\hat{\phi}^I - \phi_0\right) = &\left\{\delta - A_0(t)\exp(\beta_0^\top x)\right\} x^\top(\hat{\beta}^I - \beta_0) \\ &+ \left\{\nabla_{\theta_0} \log p_X(x, \theta)|_{\theta=\theta_0}\right\}^\top (\hat{\theta}^I - \theta_0) \\ &- \int_0^\infty I(u \leq t)\exp(\beta_0^\top x)\mathrm{d}(\hat{A}^I - A_0)(u).\end{aligned}$$

Define $\mathcal{S}_{\phi_0}(r, g_r) = E\left\{\mathcal{S}_{\phi_0}(Z)\middle| \mathcal{C} = r, G_{\mathcal{C}}(Z) = g_r\right\}$ similar to Tsiatis (2006, Section 7.3). Finally, let $q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}$ be the influence function of the initial estimator, $\hat{\phi}^I$, so that

$$n^{1/2}(\hat{\phi}^I - \phi_0)(t) = n^{-1/2}\sum_{i=1}^{n} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t) + o_p(1).$$

**Theorem 1** *Under the regularity conditions in the appendix,*

$$\left[n^{1/2}\left(\hat{\beta} - \beta_0\right), n^{1/2}\left\{\hat{A}(t) - A_0(t)\right\}_{t\in[0,\tau]}\right]$$

*converges in distribution to a tight mean zero Gaussian process in $\mathbb{R}^p \times \ell^\infty[0, \tau]$. In particular,*

$$n^{1/2}(\hat{\beta} - \beta_0) \to N\left\{0, (I^F)^{-1}\Sigma\,(I^F)^{-1}\right\} \tag{3}$$

*in distribution, where*

$$\begin{aligned}\Sigma = m^{-1}E&\left[\mathrm{var}\left\{S^F_{\mathrm{eff}}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right]\\ &+\mathrm{var}\left[E\left\{S^F_{\mathrm{eff}}(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\right\} + D_{\mathrm{eff}}(\phi_0)q\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\end{aligned}$$

*and $D_{\mathrm{eff}}(\phi_0) = E\left(S^F_{\mathrm{eff}}(Z)\left[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\right)$ and $I^F$ denotes the variance of the Cox partial likelihood score, i.e., the full-data information matrix for $\beta_0$.*

***Remark 1*** We omit giving an expression for the asymptotic variance of $\hat{A}$ to keep the presentation brief. In the next section, we present consistent estimators of the variance of both $\hat{\beta}$ and $\hat{A}(t)$.

***Remark 2*** Having a joint asymptotic distribution for $\hat{\beta}$ and $\hat{A}$ allows us to draw inference also about, e.g., the survival function $S_0(t, x) = \exp\{-A_0(t)\exp(\beta_0^\top x)\}$ for a subject with covariates $x$. To do so, we may use that

$$\begin{aligned}&n^{1/2}\{\hat{S}(t, x) - S_0(t, x)\}\\ &\quad= - S_0(t, x)\exp(\beta_0^\top x)\left[n^{1/2}\{\hat{A}(t) - A_0(t)\} + A_0(t)n^{1/2}(\hat{\beta} - \beta_0)\right] + o_P(1)\end{aligned}$$

(see e.g., Andersen et al. 1992).

As mentioned in the introduction, the standard asymptotic results for multiple imputation estimators rely on empirical process tools. In particular, we would need stochastic equicontinuity of the empirical process based on $m^{-1}\sum_{j=1}^{m} S^F_{\mathrm{eff}}\{Z_{ij}(\phi)\}$. However, as we show in the appendix, if any of the missing explanatory variables are categorical, this process is not stochastic equicontinuous. To circumvent this problem, we split $m^{-1}\sum_{j=1}^{m} S^F_{\mathrm{eff}}\{Z_{ij}(\phi)\}$ in $m^{-1}\sum_{j=1}^{m}(S^F_{\mathrm{eff}}\{Z_{ij}(\phi)\} - E[S^F_{\mathrm{eff}}\{Z_{ij}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)])$

and $\sum_{j=1}^{m} E[S_{\text{eff}}^{F}\{Z_{ij}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]$, show that the empirical process corresponding to the latter term is stochastic equicontinuous and handle the former term using a conditional central limit theorem conditioning on the observed data. Further details are given in the appendix.

## 4 Estimation of the variance

The variance of a multiple imputation estimator is usually estimated by combining of the complete data variance estimators and an estimate of the between imputation variance. For the regression parameters of the Cox model, the variance would be estimated by:

$$\left(\hat{I}^F\right)^{-1} + \left(1 + \frac{1}{m}\right) \frac{1}{m-1} \sum_{j=1}^{m} (\hat{\beta}_j - \hat{\beta})^2 \tag{4}$$

where $\hat{I}^F = \frac{1}{m} \sum_{j=1}^{m} \hat{I}_j^F$, with

$$\hat{I}_j^F = n^{-1} \sum_{i=1}^{n} \left( \frac{\sum_{l=1}^{n} S_2\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}}{\sum_{l=1}^{n} S_0\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}} - \left[ \frac{\sum_{l=1}^{n} S_1\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}}{\sum_{l=1}^{n} S_0\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}} \right]^{\otimes 2} \right) \Delta_i,$$

the full-data observed information matrix from Cox's partial likelihood based on the imputed data. It is, however, generally accepted that the validity of this estimator relies on the imputations being at least approximately drawn from Bayesian predictive distribution (Rubin 1996; Wang and Robins 1998). Using the results we derive in the appendix, it is easily seen that (4) fails to estimate the variance of $\hat{\beta}$, and the simulations in Sect. 6 indicate that (4) underestimates the variance in line with Tsiatis (2006, p. 365).

In the following, we will derive a consistent variance estimator. The multiple-imputation estimator of $\beta$ is not asymptotically linear in general as the imputations are not generally sufficiently "smooth" as functions of the initial estimator. It does, however, have the same asymptotic distribution as $n^{-1/2} \sum_{i=1}^{n} (I^F)^{-1} \xi_i$, where

$$\xi_i = \frac{1}{m} \sum_{j=1}^{m} S_{\text{eff}}^{F}\{Z_{ij}(\phi_0)\} + D_{\text{eff}}\, q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}, \tag{5}$$

by the central limit theorem as $var(\xi_1) = \Sigma$. Hence, the variance of the estimator of $\beta_0$ can be estimated consistently by

$$\left(\hat{I}^F\right)^{-1} n^{-1} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^{\top} \left(\hat{I}^F\right)^{-1},$$

if we can provide reasonable estimates, $\hat{\xi}_i$, of $\xi_i$, so that $n^{-1}\sum_{i=1}^n \hat{\xi}_i\hat{\xi}_i^\top$ is a consistent estimator of $\Sigma$. It follows from lemma 4 in the appendix that $\hat{I}^F$ is a consistent estimator of the full-data expected information $I^F$.

To obtain $\hat{\xi}_i$ we first note that as shown in the appendix we can replace the efficient score, $S_{\mathrm{eff}}^F\{Z_{ij}(\phi_0)\}$, in (5) involving the infeasible perfect imputations $Z_{ij}(\phi_0)$ by the Cox partial score function with the actual imputations $Z_{ij}(\hat{\phi}^I)$.

Next, to estimate the term $D_{\mathrm{eff}}(\phi_0)q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$, we first replace $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ by its empirical counter part $\hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}$ (cf. assumption 7 in the appendix). To estimate $D_{\mathrm{eff}}(\phi_0)$, we need to estimate

$$E\left\{S_{\mathrm{eff}}^F(Z)S_\eta{}^\top(Z,\phi_0)\right\} - E\left[S_{\mathrm{eff}}^F(Z)S_\eta{}^\top\{\mathcal{C}, G_{\mathcal{C}}(Z),\phi_0\}\right]$$

where $\eta = (\beta,\theta)$ and $S_\eta(z,\phi) = \partial/\partial\eta \log \tilde{p}_Z(z,\phi)$ is the score for $\eta$, as well as the mean in the integral

$$\int_0^\tau E\left(S_{\mathrm{eff}}^F(Z)I(u\le T)\left[\exp(\beta_0^\top X) - E\left\{\exp(\beta_0^\top X)\middle|\, \mathcal{C}, G_{\mathcal{C}}(Z)\right\}\right]\right) \\ \times\, \mathrm{d}\hat{q}_A\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(u) \tag{6}$$

where $\hat{q}_A$ is the part of $\hat{q}$ corresponding to $A$. Hence, we need to estimate means of the form $E\left\{S_{\mathrm{eff}}^F(Z)f(Z,\phi_0,u)\right\}$ and $E\left[S_{\mathrm{eff}}^F(Z)E\{f(Z,\phi_0,u)|\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]$ for suitable functions $f$. The first type of terms can be estimated consistently by

$$m^{-1}\sum_{j=1}^m n^{-1}\sum_{i=1}^n\left[X_{ij}(\hat{\phi}^I) - \frac{\sum_{l=1}^n S_1\{T_i, Z_{lj}(\hat{\phi}^I),\hat{\beta}_j\}}{\sum_{l=1}^n S_0\{T_i, Z_{lj}(\hat{\phi}^I),\hat{\beta}_j\}}\right]\Delta_i f\{Z_{ij}(\hat{\phi}^I),\hat{\phi}_j,u\},$$

where $\hat{\phi}_j = (\hat{\beta}_j, \hat{A}_j, \hat{\theta}^I)$, using corollary 1 and lemma 4. The second type of terms can be estimated consistently by

$$n^{-1}\sum_{i=1}^n\{m(m-1)\}^{-1}\sum_{\substack{j,j'=1\\ j\ne j'}}^m\left[X_{ij}(\hat{\phi}^I) - \frac{\sum_{l=1}^n S_1\{T_i, Z_{lj}(\hat{\phi}^I),\hat{\beta}_j\}}{\sum_{l=1}^n S_0\{T_i, Z_{lj}(\hat{\phi}^I),\hat{\beta}_j\}}\right] \\ \times\,\Delta_i f\{Z_{ij'}(\hat{\phi}^I),\hat{\phi}_{j'},u\},$$

as

$$\begin{aligned} &E\left[S_{\mathrm{eff}}^F(Z)E\{f(Z,\phi_0,u)|\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] \\ &= E\left[E\{S_{\mathrm{eff}}^F(Z)|\mathcal{C}, G_{\mathcal{C}}(Z)\}E\{f(Z,\phi_0,u)|\mathcal{C}, G_{\mathcal{C}}(Z)\}\right] \\ &= E\left[S_{\mathrm{eff}}^F\{Z_{ij}(\phi_0)\}f\{Z_{ij'}(\phi_0),\phi_0,u\}\right] \end{aligned}$$

for $j \neq j'$. This allows us to estimate the means that form $D_{\text{eff}}(\phi_0)$ giving us the estimator $\widehat{D_{\text{eff}}}$.

We summarize this as

**Theorem 2** *The asymptotic variance of the multiple imputation estimator $\hat{\beta}$ can be estimated by*

$$\left(\hat{I}^F\right)^{-1} n^{-1} \sum_{i=1}^{n} \hat{\xi}_i \hat{\xi}_i^\top \left(\hat{I}^F\right)^{-1}$$

*where $\hat{I}^F$ is the average of the observed data information matrices from Cox's partial likelihood based on the $m$ sets of imputations and*

$$\hat{\xi}_i = \frac{1}{m} \sum_{j=1}^{m} \left[ X_{ij}(\hat{\phi}^I) - \frac{\sum_{l=1}^{n} S_1\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}}{\sum_{l=1}^{n} S_0\{T_i, Z_{lj}(\hat{\phi}^I), \hat{\beta}_j\}} \right] \Delta_i + \widehat{D_{\text{eff}}} \hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\},$$

*with $\widehat{D_{\text{eff}}}$ as described above.*

Estimation of the variance of $\hat{A}(t)$ may be done in a similar manner: $n^{1/2}\{\hat{A}(t) - A_0(t)\}$ has the same asymptotic distribution as

$$n^{-1/2} \sum_{i=1}^{n} \rho_i^A(t), \tag{7}$$

where

$$\begin{aligned} \rho_i^A(t) = &- \int_0^t \frac{s_1(u)}{s_0(u)} \alpha_0(u) \mathrm{d}u \, (I^F)^{-1} \xi_i + m^{-1} \sum_{j=1}^{m} \int_0^t \frac{\mathrm{d}M^F\{u, Z_{ij}(\phi_0)\}}{s_0(u)} \\ &- \left\{ \int_0^t \frac{E\left(S_0(u, Z)\left[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\right)}{s_0(u)} \alpha_0(u) \mathrm{d}u \right\} q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} \end{aligned}$$

and its variance can be estimated by $n^{-1} \sum_{i=1}^{n} \left\{\hat{\rho}_i^A(t)\right\}^2$, where $\hat{\rho}_i^A$ approximates $\rho_i^A$ using techniques parallel to what was outlined for $\beta$ above.

## 5 Iterated multiple imputation

The efficiency of the multiple-imputation estimator depends on the number of imputations and on the efficiency of the initial estimator. Clearly, the efficiency increases with the number of imputations, but we would also expect the multiple imputation estimator with a sufficiently large number of imputations to improve on an inefficient initial estimator. Obviously, if the initial estimator is fully efficient, imputing the missing data will not improve the estimation. If on the other hand the initial estimator is the complete-case estimator (if the data is missing completely at random) or a simple inverse probability of missingness weighted estimator, imputation will allow us to use the incomplete observations, too. Indeed, for both of these initial estimators, the estimator of the integrated hazard will only jump at event times for which we have complete data, whereas the multiple-imputation estimator will jump whenever we observe an event time allowing the multiple imputation estimator to better approximate the unknown smooth integrated baseline hazard. An obvious idea for how to improve the estimation further would be to iterate the imputation: First estimate the parameters using multiple imputations based on an inefficient initial estimator. It will typically be beneficial to re-estimate $\theta$ as well based on the multiply imputed data. Then generate new imputations based on the multiple-imputation estimator and estimate the unknown parameters again. Obviously, this iteration scheme may be repeated several times. The final estimator is again a multiple imputation estimator based on an initial estimator which is now a multiple imputation estimator. Unfortunately, the proofs of our asymptotic results rely on the initial estimator being asymptotically linear, which the imputation estimator is not guaranteed to be as that requires stochastic equicontinuity. Hence, a new argument, which we outline in the appendix, is required to secure the asymptotic results for the iterated estimator. The conclusion is that the iterated multiple-imputation estimator is asymptotically Gaussian and that its variance may be estimated as outlined in the previous section with $\hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)$ replaced by $\hat{\rho}_i(t) = \{\hat{\rho}_i^{\beta}, \hat{\rho}_i^{A}(t), \hat{\rho}_i^{\theta}\}$ where $\hat{\rho}_i^{\beta} = (I^F)^{-1}\hat{\xi}_i$, $\hat{\xi}_i$ and $\hat{\rho}_i^{A}$ were defined in Sect. 4, and $\hat{\rho}_i^{\theta}$ is an estimate of the influence function of the multiple imputation estimator for $\theta$ obtained using techniques similar to those used to get $\hat{\rho}_i^{\beta}$.

## 6 Simulation study

We simulated covariates $X_3 \sim N(1, 0.5)$, $X_2 \sim \text{Bernoulli}\{p = \text{expit}(-1 + 0.5X_3)\}$, $X_1 \sim N(-0.25 + X_2 - 0.5X_3, 1)$. Event times were generated from the hazard $\alpha(t) = \lambda\nu t^{\nu-1}\exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3)$, with Weibull baseline parameters $\nu = 0.5$, $\lambda = 0.1$ and regression coefficients $\beta_1 = -0.2$, $\beta_2 = 0.3$, $\beta_3 = 0.5$, $\beta_4 = 0.2$. Right-censoring times were generated from an exponential distribution with mean 100. Durations longer than $\tau = 100$ were right-censored.

The covariates $X_1$ and $X_2$ were missing at random according to three different missing data mechanisms. In the first, $pr(X_1 \text{ missing}|X_3) = \text{expit}(-3 + X_3)$ and $pr(X_2 \text{ missing}|X_3) = \text{expit}(1 - 2X_3)$, leading to approximately 60% complete cases. In the second, $pr(X_1 \text{ missing}|X_3) = \text{expit}(-1 + X_3)$ and $pr(X_2 \text{ missing}|X_3) = \text{expit}(2 - 2X_3)$, leading to approximately 23% complete cases. In the last, $pr(X_1 \text{ missing}|X_3) = \text{expit}(0.6X_3)$ and $pr(X_2 \text{ missing}|X_3) = \text{expit}(1.6 - X_3)$, leading to approximately 12% complete cases. For the first scenario, we used a moderate sample size ($n = 500$), while for the latter two, we used a larger sample size ($n = 2000$).

Table 1 summarizes 10,000 repeated simulations with $m = 20$ imputations and using rejection sampling as in Bartlett et al. (2015) to generate the imputations. Increasing $m$ to 40 had no notable effect on the precision. The complete-case estimator was

**Table 1** Simulation study

| | Complete case | | | | Multiple imputation | | | | Iterated multiple imputation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | ESE | ASE | ECP | Bias | ESE | ASE | ECP | Bias | ESE | ASE | ECP |
| $n = 500$, $X_1$ *missing for* 13% *and* $X_2$ *missing for* 30%, 60% *complete cases* | | | | | | | | | | | | |
| $\beta_1$ | 0.4 | 6.9 | 6.8 | 94.3 | 0.4 | 6.1 | 5.9 | 94.4 | 0.4 | 6.1 | 5.9 | 94.4 |
| $\beta_2$ | 0.4 | 39.7 | 38.4 | 94.2 | − 0.1 | 36.1 | 34.6 | 93.9 | − 0.6 | 35.7 | 34.3 | 93.9 |
| $\beta_3$ | − 0.6 | 20.4 | 19.9 | 94.4 | − 0.5 | 15.3 | 15.0 | 94.5 | − 0.5 | 15.2 | 15.0 | 94.6 |
| $\beta_4$ | − 1.0 | 31.6 | 30.2 | 93.8 | − 0.6 | 27.4 | 26.2 | 93.8 | − 0.3 | 27.2 | 26.0 | 94.0 |
| $A(\tau/2)$ | − 1.5 | 17.8 | 17.2 | 94.5 | − 0.8 | 12.9 | 12.7 | 94.8 | − 0.7 | 12.8 | 12.7 | 94.8 |
| $A(3\tau/4)$ | − 2.0 | 21.6 | 21.2 | 94.7 | − 1.1 | 15.8 | 15.6 | 94.8 | − 1.0 | 15.7 | 15.5 | 94.6 |
| $A(\tau)$ | − 2.6 | 25.5 | 24.7 | 94.5 | − 1.4 | 18.5 | 18.1 | 94.8 | − 1.3 | 18.3 | 18.1 | 94.6 |
| $n = 2000$, $X_1$ *and* $X_2$ *missing for* 50% *each*, 23% *complete cases* | | | | | | | | | | | | |
| $\beta_1$ | 0.2 | 5.6 | 5.5 | 94.4 | 0.2 | 4.7 | 4.6 | 93.9 | 0.2 | 4.3 | 4.2 | 94.4 |
| $\beta_2$ | − 0.1 | 32.8 | 31.9 | 94.4 | − 0.5 | 25.7 | 24.9 | 94.2 | − 0.9 | 22.1 | 21.6 | 94.4 |
| $\beta_3$ | − 0.4 | 16.8 | 16.6 | 94.6 | − 0.3 | 8.6 | 8.6 | 94.9 | − 0.2 | 8.0 | 8.0 | 94.9 |
| $\beta_4$ | − 0.5 | 26.0 | 25.3 | 94.2 | 0.0 | 17.8 | 17.4 | 94.3 | 0.2 | 15.4 | 15.2 | 94.7 |
| $A(\tau/2)$ | − 1.1 | 14.5 | 14.2 | 94.8 | − 0.3 | 7.7 | 7.6 | 95.0 | − 0.1 | 7.0 | 7.0 | 94.9 |
| $A(3\tau/4)$ | − 1.5 | 17.8 | 17.5 | 94.8 | − 0.4 | 9.4 | 9.3 | 94.9 | − 0.2 | 8.5 | 8.5 | 94.9 |
| $A(\tau)$ | − 1.7 | 20.6 | 20.4 | 94.8 | − 0.5 | 10.8 | 10.7 | 95.2 | − 0.3 | 9.9 | 9.8 | 95.1 |

**Table 1** continued

| | Complete case | | | | Multiple imputation | | | | Iterated multiple imputation | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bias | ESE | ASE | ECP | Bias | ESE | ASE | ECP | Bias | ESE | ASE | ECP |
| $n = 2000$, $X_1$ *and* $X_2$ *missing for* 64% *each*, 12% *complete cases* | | | | | | | | | | | | |
| $\beta_1$ | 0.4 | 7.9 | 7.6 | 93.9 | 0.3 | 6.6 | 6.2 | 93.6 | 0.4 | 5.3 | 5.1 | 94.0 |
| $\beta_2$ | 0.3 | 41.9 | 40.5 | 94.1 | − 0.2 | 32.5 | 31.3 | 93.6 | − 0.6 | 24.9 | 24.2 | 94.2 |
| $\beta_3$ | − 0.9 | 21.9 | 21.1 | 94.0 | − 0.4 | 10.4 | 10.2 | 94.9 | − 0.2 | 8.6 | 8.6 | 95.1 |
| $\beta_4$ | − 0.9 | 33.7 | 32.3 | 93.8 | − 0.1 | 23.3 | 22.3 | 93.5 | − 0.2 | 17.3 | 16.9 | 94.3 |
| $A(\tau/2)$ | − 1.6 | 18.6 | 17.9 | 94.5 | − 0.4 | 9.1 | 8.9 | 94.9 | − 0.2 | 7.5 | 7.5 | 94.8 |
| $A(3\tau/4)$ | − 2.2 | 23.0 | 22.0 | 94.4 | − 0.7 | 11.0 | 10.8 | 94.9 | − 0.4 | 9.1 | 9.1 | 95.0 |
| $A(\tau)$ | − 2.7 | 26.8 | 25.7 | 94.5 | − 0.8 | 12.7 | 12.5 | 94.8 | − 0.5 | 10.5 | 10.4 | 95.1 |

All values have been multiplied by 100

*ESE* Empirical standard error; *ASE* average of the estimated standard error, *ECP* empirical coverage probability of the 95% confidence intervals; ESE ratio used as initial estimator, as this is an asymptotically linear, unbiased estimator of the unknown parameters under the missing data mechanisms used here. The confidence intervals for the cumulative baseline were calculated using a log-transformation.

In all scenarios, the multiple-imputation estimators appear to produce unbiased estimates and yield considerably smaller standard errors compared to the complete-case estimator. In the simulations where the probability of missingness is smaller, the variance of the multiple imputation estimator is 17–47% smaller than that of the complete-case estimator. Iterating the multiple imputation estimator leads to a negligible improvement. In the simulations with larger rates of missing data, the variance of the multiple imputation estimator is 31–78% smaller than the complete-case estimator variance. Here, iterating the multiple imputation estimator leads to another 13–45% improvement.

The estimator of the cumulative baseline performs very well in terms of standard error and confidence interval coverage in all settings. The bias is small in all cases but noticeably smaller for the multiple imputation estimators. The coverage of the confidence intervals for the regression parameters is reasonable in all settings.

For completeness, we compared the average $\hat{\beta}$ variance estimates using the biased estimator given in (4) to the empirical variance. As expected, the estimator (4) underestimates the variance. In the first scenario (moderate missingness), the variance is underestimated by 5–23%, in the second scenario the variances are estimated 23–57% too low, and in the third scenario (heavy missingness), the underestimation is 44–72%.

## 7 Example: survival with liver cirrhosis

CSL1 was a double blind randomized clinical trial conducted by the Copenhagen Study Group for Liver Diseases (Schlichting et al. 1983). In the period 1962–1969, 488 patients with liver cirrhosis were treated with either the active drug prednisone (251 patients), or placebo (237 patients). The purpose of the trial was to evaluate the effect of treatment on survival after randomization. Patients were followed to either death, drop-out or end of study in September 1974. 142 prednisone patients and 150 placebo patients died during follow-up. The survival times for the remaining patients were right-censored.

The covariates recorded at entry into the trial were treatment, 0 if prednisone and 1 if placebo; sex; age at entry; antinuclear factor (an unspecific serological indicator of self-perpetuated autoimmune processes), 0 if not present and 1 if ++ to +++; and acetylcholinesterase in μmol/min/ml. Schlichting et al. (1983) found that antinuclear factor interacts with treatment. Therefore, we include this interaction in our substantive Cox model.

Antinuclear factor is missing for 153 (31%) patients and acetylcholinesterase is missing for 43 (9%) patients. Only 300 (61%) of the patients have fully observed covariate data. We assume that data are missing completely at random. Since treatment was randomized, this covariate is left out from the imputation model. It is not necessary to specify a distribution for sex and age, which have no missing values. The problem is thus reduced to the specification of the joint conditional distribution of antinuclear factor and acetylcholinesterase. We model acetylcholinesterase by linear regression on sex, age and antinuclear factor, and model the conditional distribution of antinuclear factor by a logistic regression on sex and age.

Table 2 shows the estimates from the complete-case estimator, a multiple-imputation estimator with $m = 20$ using the complete-case estimator as initial estimator, and an estimator where the multiple-imputation estimator is iterated five times. The standard error estimates obtained by the multiple-imputation estimators are smaller than those of the complete-case estimator, and iterating the multiple-imputation estimator improves the precision of the estimates further. There are only minor differences between the point estimates, but effects of treatment and its interaction with the antinuclear factor becomes significant when using the multiple imputation estimators.

# A Appendix

## A.1 Assumptions

**Assumption 1** Assume that $(\beta_0, \theta_0) \in \mathcal{B} \times \Theta$ for known compact sets $\mathcal{B} \subset \mathbb{R}^p$ and $\Theta \subset \mathbb{R}^q$, and that $A_0(t)$ is strictly increasing and continuously differentiable and that $A_0(0) = 0$.

**Assumption 2** The covariates $X$ are bounded almost surely.

**Assumption 3** Data are missing at random, $pr(\mathcal{C} = r|Z = z) = pr\{\mathcal{C} = r|G_{\mathcal{C}}(Z) = G_r(z)\}$.

**Assumption 4** The full-data information matrix, $I^F$, for $\beta$ at the true parameter value is invertible.

**Assumption 5** There is a finite maximum follow-up time $\tau > 0$, when all individuals still at risk are censored, and $pr\{Y(\tau) = 1\} = pr(T = \tau) > 0$.

**Assumption 6** The censoring distribution does not depend on $\phi_0$ and potentially missing covariates, $\alpha_U(t|x)^{1-\delta} pr(U > t|x) = \alpha_U\{t|G_{X,r}(x)\}^{1-\delta} pr\{U > t|G_{X,r}(x)\}$.

**Assumption 7** There exists a consistent (but possibly inefficient) asymptotically linear estimator $\hat{\phi}^I = \{\hat{\beta}^I, \hat{A}^I(t), \hat{\theta}^I\}$ such that $n^{1/2}(\hat{\phi}^I - \phi_0)(t) = n^{-1/2}\sum_{i=1}^n q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t) + o_P(1)$, where $q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)$ are independent processes, converges weakly to a tight Gaussian process in $\mathbb{R}^p \times \ell^\infty[0, \tau] \times \mathbb{R}^q$. Further, we assume that the variance $var\{q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)\}$ can be estimated consistently by $n^{-1}\sum_{i=1}^n \hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)\hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)^\top$ for some suitable $\hat{q}\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}(t)$.

**Assumption 8** Assume that $p_{X|\mathcal{C},G}(x|r, g, \phi)$, the conditional density of $X$ given $\mathcal{C}$ and $G_{\mathcal{C}}$ with respect to a reference measure $\nu_X$, is a Lipschitz continuous function of $\phi$ (with respect to the $L_2$-norm) in a neighborhood of $\phi_0$, with an integrable Lipschitz constant, $h(x|r, g)$ such that $\int h(x|r, g)\mathrm{d}\nu_X(x)$ is a bounded function of $(r, g)$.

**Table 2** Estimates from the liver cirrhosis study

| Method | Variable | Scoring | $\beta$ | $se(\beta)$ | $p$ |
|---|---|---|---|---|---|
| CC | Treatment | Placebo | 0.534 | 0.317 | 0.096 |
| | Sex | Male | 0.501 | 0.182 | 0.009 |
| | Age | Years | 0.036 | 0.009 | $<0.001$ |
| | Antinuclear factor (ANF) | Present | 0.305 | 0.291 | 0.231 |
| | Acetylcholinesterase | μmol/min/ml | $-0.003$ | 0.001 | $<0.001$ |
| | ANF × Treatment | Present:Placebo | $-0.675$ | 0.365 | 0.072 |
| MI | Treatment | Placebo | 0.707 | 0.277 | 0.015 |
| | Sex | Male | 0.462 | 0.148 | 0.003 |
| | Age | Years | 0.045 | 0.007 | $<0.001$ |
| | Acetylcholinesterase | Present | 0.446 | 0.285 | 0.105 |
| | Antinuclear factor (ANF) | Antinuclear factor (ANF) | $-0.003$ | 0.001 | $<0.001$ |
| | ANF × Treatment | Present:Placebo | $-0.878$ | 0.357 | 0.019 |
| Iterated MI | Treatment | Placebo | 0.672 | 0.263 | 0.015 |
| | Sex | Male | 0.444 | 0.151 | 0.005 |
| | Age | Years | 0.046 | 0.007 | $<0.001$ |
| | Antinuclear factor (ANF) | Present | 0.476 | 0.279 | 0.094 |
| | Acetylcholinesterase | μmol/min/ml | $-0.003$ | 0.001 | $<0.001$ |
| | ANF × Treatment | Present:Placebo | $-0.841$ | 0.327 | 0.014 |

*CC* Complete case, *MI* multiple imputation, *SE* standard error; $p$ $p$-value

### A.2 Lemmas

We first introduce some notation. The density of the (potentially unobserved) full data $z = (t, \delta, x)$ and the observed data $\{r, g = (t, \delta, g_x)\}$ are

$$
\begin{aligned}
p_{\mathcal{C},Z}(r, z, \phi) &= pr(\mathcal{C} = r | Z = z)\alpha_U(t|x)^{1-\delta} pr(U > t|x) \\
&\quad \times \alpha(t)^\delta \exp\left\{\delta\beta^\top x - A(t)\exp(\beta^\top x)\right\} p_X(x, \theta) \\
&= pr\{\mathcal{C} = r | G_{\mathcal{C}}(Z) = G_r(z)\}\alpha_U\{t|G_{X,r}(x)\}^{1-\delta} \\
&\quad \times pr\{U > t|G_{X,r}(x)\}\alpha(t)^\delta \exp\left\{\delta\beta^\top x - A(t)\exp(\beta^\top x)\right\} p_X(x, \theta), \\
p_{\mathcal{C},G}(r, g, \phi) &= \int_{\{G_r(z)=g\}} p_{\mathcal{C},Z}(r, z, \phi) d\nu_Z(z) \\
&= pr(\mathcal{C} = r | G_{\mathcal{C}}(Z) = g)\alpha_U\{t|G_{X,r}(x)\}^{1-\delta} pr\{U > t|G_{X,r}(x)\} \\
&\quad \times \alpha(t)^\delta \int_{\{G_{X,r}(x)=g_x\}} \exp\left\{\delta\beta^\top x - A(t)\exp(\beta^\top x)\right\} p_X(x, \theta) d\nu_X(x),
\end{aligned}
$$

where $\nu.(\cdot)$ is a dominating measure for which the densities of the random variables are defined. Recall the definition $\tilde{p}_Z(z, \phi) = \exp\left\{\delta\beta^\top x - A(t)\exp\left(\beta^\top x\right)\right\} p_X(x, \theta)$ and let $\tilde{p}_G(g, \phi) = \int_{\{G_r(v)=g\}} \tilde{p}_Z(v) \mathrm{d}\nu_Z(v)$. Note that

$$
\frac{p_{\mathcal{C},Z}(r, z, \phi)}{p_{\mathcal{C},G}\{r, G_r(z), \phi\}} = \frac{\tilde{p}_Z(z, \phi)}{\tilde{p}_G\{G_r(z, \phi)\}}.
$$

The following lemma building on Wang and Robins (1998), Robins and Wang (2000), see also Tsiatis (2006, Lemma 14.2), will be used repeatedly.

**Lemma 1** *For $f(t, Z)$, continuous in $t \in [0, \tau]$ and bounded with probability one,*

$$
\begin{aligned}
&n^{1/2} E\left[f\{t, Z(\phi)\} - f\{t, Z(\phi_0)\}\right]_{|\phi=\hat{\phi}^I} \\
&\quad = E\left(f(t, Z)\left[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}\right]\right) n^{1/2}(\hat{\phi}^I - \phi_0) + o_p(1)
\end{aligned}
$$

*where the remainder term is uniform in $t$.*

***Proof*** Following Tsiatis (2006, pp. 350–352), we write

$$
\begin{aligned}
E\left[f\{t, Z(\phi)\}\right] &= E\left(E\left[f\{t, Z(\phi)\}| \mathcal{C}, G_{\mathcal{C}}(Z), \phi\right]\right) \\
&= \int f(t, z)\frac{p_{\mathcal{C},Z}(r, z, \phi)}{p_{\mathcal{C},G}\{r, G_r(z), \phi\}} p_{\mathcal{C},G}\{r, G_r(z), \phi_0\}\mathrm{d}\nu_{\mathcal{C},Z}(r, z) \\
&= \int f(t, z)\frac{\tilde{p}_Z(z, \phi)}{\tilde{p}_G\{G_r(z), \phi\}} p_{\mathcal{C},G}\{r, G_r(z), \phi_0\}\mathrm{d}\nu_{\mathcal{C},Z}(r, z)
\end{aligned}
$$

so that

$$
E\left[f\{t, Z(\phi)\} - f\{t, Z(\phi_0)\}\right]_{|\phi=\hat{\phi}^I}
$$

$$
\begin{aligned}
&= \int f(t,z)\left[\frac{\tilde p_Z(z,\hat\phi^I)}{\tilde p_G\{G_r(z),\hat\phi^I\}} - \frac{\tilde p_Z(z,\phi_0)}{\tilde p_G\{G_r(z),\phi_0\}}\right] p_{\mathcal{C},G}\{r,G_r(z),\phi_0\}\mathrm{d}\nu_{\mathcal{C},Z}(r,z) \\
&= \int f(t,z)\frac{\tilde p_Z(z,\phi_0)}{\tilde p_G\{G_r(z),\phi_0\}}\left[\mathcal{S}_{\phi_0}(z) - \mathcal{S}_{\phi_0}\{r,G_r(z)\}\right](\hat\phi^I - \phi_0) \\
&\quad \times p_{\mathcal{C},G}\{r,G_r(z),\phi_0\}\mathrm{d}\nu_{\mathcal{C},Z}(r,z) + o_P(n^{-1/2}) \\
&= \int f(t,z)\left[\mathcal{S}_{\phi_0}(z) - \mathcal{S}_{\phi_0}\{r,G_r(z)\}\right](\hat\phi^I - \phi_0) p_{\mathcal{C},Z}(r,z,\phi_0)\mathrm{d}\nu_{\mathcal{C},Z}(r,z) \\
&\quad + o_P(n^{-1/2}) \\
&= E\left(f(t,Z)\left[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C},G_{\mathcal{C}}(Z)\}\right]\right)(\hat\phi^I - \phi_0) + o_P(n^{-1/2}).
\end{aligned}
$$

□

**Lemma 2** *Let $f[\{X_{ij}(\phi),\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}_{j=1,\ldots,m}]$ be a bounded function. Then the logarithm of the $\epsilon$-bracketing number of the class*

$$
\begin{aligned}
&\{(r,g)\mapsto E(f[\{X_{ij}(\phi)\}_{j=1,\ldots,m},\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)]|\mathcal{C}_i = r, G_{\mathcal{C}_i}(Z_i) = g) \\
&\quad : \|\phi - \phi_0\|_{L_2} \le \delta\}
\end{aligned}
\tag{8}
$$

*is bounded by a constant times $1/\epsilon$.*

***Proof*** Let $F_i(\phi) = E(f[\{X_{ij}(\phi)\}_{j=1,\ldots,m},\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)]|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i))$. Then

$$
\begin{aligned}
&|F_i(\phi) - F_i(\phi_0)| \\
&\quad \le \int |f\{x,\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}||p_{X|\mathcal{C},G}\{x|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\phi\} \\
&\qquad - p_{X|\mathcal{C},G}\{x|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i),\phi_0\}|d\nu_X(x) \\
&\quad \le \mathit{constant} \times \|\phi - \phi_0\|_{L_2}
\end{aligned}
$$

by assumption 8. It follows that the bracketing number of the class (8) is bounded by the bracketing number of $\{\phi : \|\phi - \phi_0\|_{L_2} \le \delta\}$ and this is dominated by the bracketing number of the integrated baseline hazard which is smaller than $\exp(K/\epsilon)$ by van der Vaart and Wellner (1996, Theorem 2.7.5) for a constant $K$. □

It follows that for a bounded function $f$, the process

$$
n^{-1/2}\sum_{i=1}^{n}\left\{E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}]|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)) - E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}])\right\}
$$

is stochastic equicontinuous near $\phi_0$, and that

$$
n^{-1}\sum_{i=1}^{n} E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}]|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i))
$$

converges almost surely, uniformly in a neighborhood of $\phi_0$. The process

$$n^{-1/2}\sum_{i=1}^{n}\left\{f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}] - E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}])\right\}$$

is not stochastic equicontinuous in general. A proof of this is included at the end of this appendix.

We will need some results for averages of functions of the imputations and the unknown parameter.

**Lemma 3** *Let $f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\phi]$ be a bounded function which is Lipschitz continuous as a function of $\phi$ in a neighborhood of $\phi_0$ with a bounded Lipschitz constant. Then*

$$n^{-1}\sum_{i=1}^{n} f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\tilde{\phi}] - E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m},\phi_0]|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i))_{|\phi=\hat{\phi}^I}$$

*converges to in probability to 0 for any consistent estimator $\tilde{\phi}$ of $\phi_0$.*

***Proof*** As $|f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\tilde{\phi}] - f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\phi_0]| \le \textit{constant} \times \|\tilde{\phi} - \phi_0\|_{L_2}$, we only need to consider the case where $\tilde{\phi} = \phi_0$. Letting

$$F_i = f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\phi_0] - E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m},\phi_0]|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i))_{|\phi=\hat{\phi}^I}$$

we see that $E\{F_i|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} = 0$ so that

$$\mathrm{var}\left(n^{-1}\sum_{i=1}^{n}F_i\right) = E\left[n^{-2}\sum_{i=1}^{n}\mathrm{var}\left\{F_i|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right\}\right] = O(1/n)$$

as $F_i$ is bounded by assumption. □

**Corollary 1** *Let $f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\phi]$ be a bounded function, which is Lipschitz continuous as a function of $\phi$ in a neighborhood of $\phi_0$ with a bounded Lipschitz constant. Suppose further that $E(f[\{Z_{ij}(\phi')\}_{j=1,\ldots,m},\phi_0])$ is a continuous function of $\phi'$. Then*

$$n^{-1}\sum_{i=1}^{n} f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\tilde{\phi}] \to E(f[\{Z_{1j}\}_{j=1,\ldots,m},\phi_0])$$

*in probability for any consistent estimator $\tilde{\phi}$ of $\phi_0$.*

***Proof*** The average $n^{-1}\sum_{i=1}^{n} f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\tilde{\phi}]$ may be split into a sum of

$$n^{-1}\sum_{i=1}^{n} f[\{Z_{ij}(\hat{\phi}^I)\}_{j=1,\ldots,m},\tilde{\phi}] - E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m},\phi_0]|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i))_{|\phi=\hat{\phi}^I}$$

which is $o_P(1)$ by lemma 3, and $n^{-1}\sum_{i=1}^n E(f[\{Z_{ij}(\phi)\}_{j=1,\ldots,m}, \phi_0]|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i))_{|\phi=\hat{\phi}^I}$ which converges to $E(f[\{Z_{1j}\}_{j=1,\ldots,m}, \phi_0])$ by lemma 2 and the uniform law of large numbers. □

**Lemma 4** *If $\tilde{\beta} \to \beta_0$ in probability, then*

$$n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \tilde{\beta}\} \to s_k(t) \qquad (k = 0, 1, 2,\ j = 1, \ldots, m)$$

*in probability, uniformly in $t \in [0, \tau]$.*

***Proof*** It suffices to consider the case where $X$ is one-dimensional. Clearly, by differentiability and boundedness,

$$\sup_{t\in[0,\tau]}\left|n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \tilde{\beta}\} - n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \beta_0\}\right| \leq \textit{constant} \times |\tilde{\beta} - \beta_0|$$

so we may replace $\tilde{\beta}$ by $\beta_0$. Furthermore, by corollary 1, $n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \beta_0\} - s_k(t) = o_P(1)$ for any $t$. Assume for simplicity $X_1 \geq 0$ with probability 1. Choose finitely many $0 = t_0 < t_1 < \cdots < t_L = \tau$ such that for any $t$ there is an $\ell$ such that $E\{Y_1(t) - Y_1(t_\ell)\}$, $E\{Y_1(t_{\ell-1}) - Y_1(t)\} \leq \epsilon/c_k$, where $c_k$ is an upper bound on $X_1^k \exp(\beta_0^\top X_1)$. Then

$$\begin{aligned} &n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \beta_0\} - s_k(t) \\ &\quad\leq n^{-1}\sum_{i=1}^n S_k\{t_{\ell-1}, Z_{ij}(\hat{\phi}^I), \beta_0\} - s_k(t_{\ell-1}) + s_k(t_{\ell-1}) - s_k(t) \leq o_P(1) + \epsilon \end{aligned}$$

where the $o_P(1)$-term does not depend on $t$. Combined with a similar lower bound, this yields the desired uniform convergence. If $pr(X_1 < 0) > 0$ we may split (when $k = 1$) $X_{ij}(\hat{\phi}^I)$ into a sum of $X_{ij}(\hat{\phi}^I) - \min X_1$ and $\min X_1$, where $\min X_1$ denotes the lower bound for the support of $X_1$ (the essential infimum). Thus, $n^{-1}\sum_{i=1}^n S_k\{t, Z_{ij}(\hat{\phi}^I), \beta_0\}$ may be split into a sum of two terms, each of which may be handled as indicated above. □

### A.3 Proof of Theorem 1: Regression parameters

The multiple-imputation estimator of $\beta_0$ is $\hat{\beta} = m^{-1}\sum_{j=1}^m \hat{\beta}_j$, where the $j$th imputation estimator $\hat{\beta}_j$ is the solution to $U_j(\hat{\beta}_j, \hat{\phi}^I) = 0$, with

$$U_j(\beta, \hat{\phi}^I) = \sum_{i=1}^n \left[X_{ij}(\hat{\phi}^I) - \frac{\sum_{l=1}^n S_1\{T_i, Z_{lj}(\hat{\phi}^I), \beta\}}{\sum_{l=1}^n S_0\{T_i, Z_{lj}(\hat{\phi}^I), \beta\}}\right]\Delta_i.$$

Following standard arguments and using lemma 4, $\hat{\beta}_j$ may be shown to be consistent and $n^{1/2}(\hat{\beta}_j - \beta_0) = n^{-1/2}\left(I^F\right)^{-1} U_j(\beta_0, \hat{\phi}^I) + o_P(1)$, where $I^F$ is the full-data information matrix for $\beta$. Averaging the $m$ estimators we get

$$n^{1/2}(\hat{\beta} - \beta_0) = n^{-1/2}\left(I^F\right)^{-1} m^{-1}\sum_{j=1}^{m} U_j(\beta_0, \hat{\phi}^I) + o_P(1). \tag{9}$$

As the imputations depend on the initial estimator, $\hat{\phi}^I$, which involves information from all subjects, this is not a sum of independent and identically distributed terms. We can write

$$\begin{aligned} n^{-1/2}U_j(\beta_0, \hat{\phi}^I) &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left[X_{ij}(\hat{\phi}^I) - \frac{\sum_{l=1}^{n} S_1\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}{\sum_{l=1}^{n} S_0\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}\right]\mathrm{d}M^F\{u, Z_{ij}(\hat{\phi}^I)\} \\ &= n^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}\left\{X_{ij}(\hat{\phi}^I) - e(u)\right\}\mathrm{d}M^F\{u, Z_{ij}(\hat{\phi}^I)\} \\ &\quad + \int_0^{\tau}\left[e(u) - \frac{\sum_{l=1}^{n} S_1\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}{\sum_{l=1}^{n} S_0\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}\right] n^{-1/2}\sum_{i=1}^{n}\mathrm{d}M^F\{u, Z_{ij}(\phi_0)\} \\ &\quad - \int_0^{\tau}\left[e(u) - \frac{\sum_{l=1}^{n} S_1\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}{\sum_{l=1}^{n} S_0\{u, Z_{lj}(\hat{\phi}^I), \beta_0\}}\right] \\ &\quad \times n^{-1/2}\sum_{i=1}^{n} Y_i(u)\left[\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\} - \exp\left\{\beta_0^\top X_{ij}(\phi_0)\right\}\right]\alpha_0(u)du. \end{aligned} \tag{10}$$

The second term on the right-hand side above converges to zero in probability by Lemma 4 and Kosorok (2008, Lemma 4.2). To show that the third term also converges to zero in probability, it suffices (by Kosorok 2008, Lemma 4.2) to show that the second factor in the integrand of (10),

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n} Y_i(u)\left[\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\} - \exp\left\{\beta_0^\top X_{ij}(\phi_0)\right\}\right] \\ &\quad = n^{-1/2}\sum_{i=1}^{n} Y_i(u)\left(\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\} - E\left[\exp\left\{\beta_0^\top X_{ij}(\phi)\right\}\middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right) \\ &\quad - n^{-1/2}\sum_{i=1}^{n} Y_i(u)\left(\exp\left\{\beta_0^\top X_{ij}(\phi_0)\right\} - E\left[\exp\left\{\beta_0^\top X_{ij}(\phi_0)\right\}\middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\right) \\ &\quad + n^{-1/2}\sum_{i=1}^{n} Y_i(u)\left(E\left[\exp\{\beta_0^\top X_{ij}(\phi)\}\middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right. \\ &\quad \left. - E\left[\exp\{\beta_0^\top X_{ij}(\phi_0)\}\middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]\right) \end{aligned} \tag{11}$$

is bounded in probability. The first two terms have mean zero and finite variance and are thus bounded in probability. By stochastic equicontinuity, continuity of the mean and $n^{1/2}$-consistency of the initial estimator, the third term is also bounded in probability. Thus,

$$
\begin{aligned}
&n^{-1/2}m^{-1}\sum_{j=1}^{m} U_j(\beta_0,\hat{\phi}^I)\\
&\quad= n^{-1/2}\sum_{i=1}^{n} m^{-1}\sum_{j=1}^{m} S_{\text{eff}}^F\{Z_{ij}(\hat{\phi}^I)\} + o_P(1)\\
&\quad= n^{-1/2}\sum_{i=1}^{n} m^{-1}\sum_{j=1}^{m}\left(S_{\text{eff}}^F\{Z_{ij}(\hat{\phi}^I)\} - E[S_{\text{eff}}^F\{Z_{ij}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]_{|\phi=\hat{\phi}^I}\right) \quad (12)\\
&\qquad+ n^{-1/2}\sum_{i=1}^{n}\left(E[S_{\text{eff}}^F\{Z_{i1}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]_{|\phi=\hat{\phi}^I} - E[S_{\text{eff}}^F\{Z_{i1}(\phi)\}]_{|\phi=\hat{\phi}^I}\right)\\
&\qquad- n^{-1/2}\sum_{i=1}^{n}\left(E[S_{\text{eff}}^F\{Z_{i1}(\phi_0)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)] - E[S_{\text{eff}}^F\{Z_{i1}(\phi_0)\}]\right) \quad (13)\\
&\qquad+ n^{-1/2}\sum_{i=1}^{n} E[S_{\text{eff}}^F\{Z_{i1}(\phi_0)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)] \quad (14)\\
&\qquad+ n^{1/2}\left(E[S_{\text{eff}}^F\{Z_{11}(\phi)\}]_{|\phi=\hat{\phi}^I} - E[S_{\text{eff}}^F\{Z_{11}(\phi_0)\}]\right) + o_P(1), \quad (15)
\end{aligned}
$$

where $E[S_{\text{eff}}^F\{Z_{i1}(\phi_0)\}]$ equals zero but has been included for clarity. Using lemma 1 we may write

$$
\begin{aligned}
&n^{1/2}\left(E[S_{\text{eff}}^F\{Z_{11}(\phi)\}]_{|\phi=\hat{\phi}^I} - E[S_{\text{eff}}^F\{Z_{11}(\phi_0)\}]\right)\\
&\quad= D_{\text{eff}}(\phi_0)n^{1/2}(\hat{\phi}^I - \phi_0) + o_P(1)\\
&\quad= n^{-1/2}\sum_{i=1}^{n} D_{\text{eff}}(\phi_0)q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\} + o_P(1)
\end{aligned}
$$

where $D_{\text{eff}}(\phi_0) = E(S_{\text{eff}}^F(Z)[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}])$. Thus, the last three terms—(13), (14), (15)—may be written as

$$
n^{-1/2}\sum_{i=1}^{n}\left(E[S_{\text{eff}}^F\{Z_{i1}(\phi_0)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)] + D_{\text{eff}}(\phi_0)q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right) + o_P(1)
$$

as (13) is $o_P(1)$ by the stochastic equicontinuity implied by lemma 2.

Lemma 2 (with a straightforward extension) also implies that

$$
n^{-1}\sum_{i=1}^{n} \text{var}[S_{\text{eff}}^F\{Z_{i1}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)] \to E\left(\text{var}[S_{\text{eff}}^F\{Z(\phi)\}|\mathcal{C}, G_{\mathcal{C}}(Z)]\right)
$$

almost surely, uniformly in a neighborhood of $\phi_0$. Assume for now (for simplicity) that $\hat{\phi}^I$ is strongly consistent. Then, conditionally on the observed data, for almost every realization,

$$\begin{aligned} &n^{-1/2}\sum_{i=1}^{n} m^{-1}\sum_{j=1}^{m}\left(S_{\text{eff}}^F\{Z_{ij}(\hat{\phi}^I)\} - E[S_{\text{eff}}^F\{Z_{ij}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]_{|\phi=\hat{\phi}^I}\right) \\ &\quad \to N\left\{0, m^{-1}E\left(var[S_{\text{eff}}^F\{Z(\phi_0)\}|\mathcal{C}, G_{\mathcal{C}}(Z)]\right)\right\} \end{aligned} \tag{16}$$

in distribution by the Lindeberg–Feller central limit theorem (van der Vaart 1998, Proposition 2.27). Using Schenker and Welsh (1988, Lemma 1) or Nielsen (2003, Lemma 1), it follows that (16) also holds unconditionally and that (12) is asymptotically independent of the observed data. Without strong consistency, we may for every subsequence extract a further subsequence where $\hat{\phi}^I$ converges almost surely to $\phi_0$. Thus, every subsequence has a subsequence, where (16) holds. Thus, the conditional characteristic function of the left-hand side of (16) converges almost surely along subsequences of subsequences to the characteristic function of the right-hand side of (16). This implies that the convergence holds in probability for the original sequence of characteristic functions and as the characteristic function is bounded this ensures that (16) holds unconditionally. The asymptotic distribution of $\hat{\beta}$ now follows.

### A.4 Proof of Theorem 1: Cumulative baseline hazard

The multiple-imputation estimator of the cumulative baseline hazard function is $\hat{A}(t) = m^{-1}\sum_{j=1}^{m}\hat{A}_j(t, \hat{\beta}_j)$, where

$$\hat{A}_j(t, \beta) = \int_0^t \frac{1}{\sum_{i=1}^n S_0\{u, Z_{ij}(\hat{\phi}^I), \beta\}} dN.(u)$$

is the estimator from the $j$th imputation where $N.(t) = \sum_{i=1}^n N_i(t)$. Let

$$\mathrm{d}M(t, Z_i) = \mathrm{d}N_i(t) - E\{Y_i(t)\exp(\beta_0^\top X_i)|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\alpha_0(t)\mathrm{d}t.$$

Then, $M.(t) = \sum_{i=1}^n M(t, Z_i)$ is a zero mean square-integrable martingale with respect to the observed filtration.

We may write $n^{1/2}\{\hat{A}(t) - A_0(t)\} = n^{1/2}\{\hat{A}(t) - \hat{A}_0(t)\} + n^{1/2}\{\hat{A}_0(t) - A_0(t)\}$, where

$$\hat{A}_0(t) = m^{-1}\sum_{j=1}^{m}\int_0^t \frac{1}{\sum_{i=1}^n S_0\{u, Z_{ij}(\hat{\phi}^I), \beta_0\}}\mathrm{d}N.(u).$$

Using lemma 4 and Kosorok (2008, Lemma 4.2), we have

$$
\begin{aligned}
&n^{1/2}\{\hat{A}(t)-\hat{A}_0(t)\}\\
&\quad=-m^{-1}\sum_{j=1}^{m}\int_0^t \frac{n^{-1}\sum_{i=1}^n S_1\{u,Z_{ij}(\hat{\phi}^I),\beta_0\}}{[n^{-1}\sum_{i=1}^n S_0\{u,Z_{ij}(\hat{\phi}^I),\beta_0\}]^2}n^{-1}\mathrm{d}N.(u)n^{1/2}(\hat{\beta}-\beta_0)+o_P(1)\\
&\quad=-\int_0^t \frac{s_1(u)}{s_0(u)}\alpha_0(u)\mathrm{d}u\, n^{1/2}(\hat{\beta}-\beta_0)+o_P(1).
\end{aligned}
$$

Now

$$
\begin{aligned}
&n^{1/2}\{\hat{A}_0(t)-A_0(t)\}\\
&\quad=n^{1/2}\left(m^{-1}\sum_{j=1}^{m}\int_0^t \frac{1}{\sum_{i=1}^n S_0\{u,Z_{ij}(\hat{\phi}^I),\beta_0\}}\mathrm{d}N.(u)-\int_0^t\alpha_0(u)\mathrm{d}u\right)\\
&\quad=m^{-1}\sum_{j=1}^{m}\int_0^t\left[\frac{1}{\sum_{i=1}^n S_0\{u,Z_{ij}(\hat{\phi}^I),\beta_0\}}\right.\\
&\qquad\left.-\frac{1}{\sum_{i=1}^n E\{Y_i(u)\exp(\beta_0^\top X_i)|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}}\right]n^{1/2}\mathrm{d}N.(u)\\
&\qquad+n^{1/2}\left[\int_0^t\frac{1}{\sum_{i=1}^n E\{Y_i(u)\exp(\beta_0^\top X_i)|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}}\mathrm{d}N.(u)-\int_0^t\alpha_0(u)\mathrm{d}u\right].
\end{aligned}
\tag{17}
$$

The second term of (17) may be rewritten as:

$$
\begin{aligned}
&\int_0^t\frac{1}{n^{-1}\sum_{i=1}^n E\{Y_i(u)\exp(\beta_0^\top X_i)|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\}}n^{-1/2}\mathrm{d}M.(u)+o_P(1)\\
&\quad=\int_0^t\frac{1}{s_0(u)}n^{-1/2}\mathrm{d}M.(u)+o_P(1)
\end{aligned}
$$

which converges to a Gaussian martingale. Before turning to the first term of (17), we note that

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^n Y_i(u)\left(\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}-E\left[\exp\left(\beta_0^\top X_i\right)|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right]\right)\\
&\quad=n^{-1/2}\sum_{i=1}^n Y_i(u)\left(\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}-E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right)\\
&\qquad+n^{-1/2}\sum_{i=1}^n Y_i(u)\left(E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\middle|\,\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right.\\
&\qquad\left.-E\left\{\exp\left(\beta_0^\top X_i\right)|\mathcal{C}_i,G_{\mathcal{C}_i}(Z_i)\right\}\right).
\end{aligned}
\tag{18}
$$

The second term of (18) is asymptotically equivalent to

$$n^{1/2}\left(E[S_0\{u, Z(\phi), \beta_0\}]_{|\phi=\hat{\phi}^I} - E\{S_0(u, Z, \beta_0)\}\right) = D_0(u, \phi_0)n^{1/2}(\hat{\phi}^I - \phi_0) + o_P(1)$$

where $D_0(u, \phi_0) = E(S_0(u, Z, \beta_0)[\mathcal{S}_{\phi_0}(Z) - \mathcal{S}_{\phi_0}\{\mathcal{C}, G_{\mathcal{C}}(Z)\}])$ by lemma 1. Thus, we may write the integrand of the first term of (17) as

$$\begin{aligned}
&n^{1/2}\left(\frac{1}{\sum_{i=1}^n S_0\{u, Z_{ij}(\hat{\phi}^I), \beta_0\}} - \frac{1}{\sum_{i=1}^n E\left[Y_i(u)\exp\left\{\beta_0^\top X_{i1}(\phi_0)\right\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]}\right) \\
&\quad = -\frac{n^{-3/2}\sum_{i=1}^n Y_i(u)\left(\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\} - E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\middle|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right)}{s_0(u)^2} \\
&\quad\quad - n^{-1}D_0(u, \phi_0)\frac{n^{1/2}(\hat{\phi}^I - \phi_0)}{s_0(u)^2} + o_P(1)
\end{aligned}$$

and hence the first term of (17) as

$$\begin{aligned}
&-\int_0^t n^{-1}\sum_{i=1}^n Y_i(u)\left(m^{-1}\sum_{j=1}^m \exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}\right. \\
&\quad \left.-E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\middle|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right)\frac{n^{-1/2}\mathrm{d}M.(u)}{s_0(u)^2} \\
&\quad -\int_0^t n^{-1/2}\sum_{i=1}^n Y_i(u)\left(m^{-1}\sum_{j=1}^m \exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}\right. \\
&\quad \left.-E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\middle|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right)\frac{\alpha_0(u)}{s_0(u)}du \\
&\quad -\int_0^t D_0(u, \phi_0)\frac{1}{s_0(u)^2}n^{-1}\mathrm{d}M.(u)n^{1/2}(\hat{\phi}^I - \phi_0) \\
&\quad -\int_0^t D_0(u, \phi_0)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\, n^{1/2}(\hat{\phi}^I - \phi_0) + o_P(1)
\end{aligned}$$

where the first and the third term are both $o_P(1)$ (Kosorok 2008, Lemma 4.2). Thus

$$\begin{aligned}
&n^{1/2}\{\hat{A}(t) - A_0(t)\} \\
&\quad = -\int_0^t n^{-1/2}\sum_{i=1}^n Y_i(u)\left(m^{-1}\sum_{j=1}^m \exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}\right. \\
&\quad\quad \left.-E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\middle|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u
\end{aligned}$$

$$
\begin{aligned}
&-\int_0^t D_0(u,\phi_0)\,\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\, n^{1/2}(\hat{\phi}^I-\phi_0)+\int_0^t \frac{1}{s_0(u)} n^{-1/2}\mathrm{d}M.(u)\\
&-\int_0^t \frac{s_1(u)}{s_0(u)}\alpha_0(u)\mathrm{d}u\, n^{1/2}(\hat{\beta}-\beta_0)+o_P(1) \qquad (19)
\end{aligned}
$$

where the three latter terms converge as processes. To show tightness of the first term, let $w(s,t)$ denote

$$
\begin{aligned}
&-\int_s^t n^{-1/2}\sum_{i=1}^n Y_i(u)\Bigg(m^{-1}\sum_{j=1}^m \exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}\\
&\quad -E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\Big|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\Bigg)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\\
&= -n^{-1/2}\sum_{i=1}^n\int_s^t Y_i(u)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\\
&\quad\times m^{-1}\sum_{j=1}^m\left(\exp\left\{\beta_0^\top X_{ij}(\hat{\phi}^I)\right\}-E\left[\exp\left\{\beta_0^\top X_{i1}(\phi)\right\}\Big|\,\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\right]_{|\phi=\hat{\phi}^I}\right).
\end{aligned}
$$

Then, clearly $E\{w(s,t)\}=E(E[w(s,t)|\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}_{i=1,\ldots,n})\}=0$ so that

$$
\begin{aligned}
E\{w(s,t)^2\} &= E(\mathrm{var}[w(s,t)|\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}_{i=1,\ldots,n}])\\
&= n^{-1}\sum_{i=1}^n E\Bigg(\left\{\int_s^t Y_i(u)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\right\}^2\\
&\qquad\times m^{-1}\mathrm{var}\left[\exp\left\{\beta_0^\top X_{i1}(\phi_0)\right\}\Big|\,\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}_{i=1,\ldots,n}\right]\Bigg)\\
&= O\{(t-s)^2\}
\end{aligned}
$$

implying (van der Vaart and Wellner 1996, Section 2.2.3) that also the first term of (19) is tight. Finally, we may write $n^{1/2}\{\hat{A}(t)-A_0(t)\}$ as a sum of

$$
\begin{aligned}
&n^{-1/2}\sum_{i=1}^n\Bigg\{\int_0^t \frac{1}{s_0(u)}\mathrm{d}M_i(u)-\int_0^t D_0(u,\phi_0)\frac{\alpha_0(u)}{s_0(u)}\mathrm{d}u\, q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\\
&-\int_0^t \frac{s_1(u)}{s_0(u)}\alpha_0(u)\mathrm{d}u\,(I^F)^{-1} \qquad (20)\\
&\times\left(E[S^F_{\mathrm{eff}}\{Z_{ij}(\phi_0)\}|\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]+D_{\mathrm{eff}}(\phi_0)q\{\mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)\}\right)\Bigg\}
\end{aligned}
$$

and

$$
\begin{aligned}
&- n^{-1/2} \sum_{i=1}^{n} \left\{ \int_0^t \frac{s_1(u)}{s_0(u)} \alpha_0(u) \mathrm{d}u \, (I^F)^{-1} \right. \\
&\times m^{-1} \sum_{j=1}^{m} \left( S_{\text{eff}}^F\{Z_{ij}(\hat{\phi}^I)\} - E[S_{\text{eff}}^F\{Z_{ij}(\phi)\} | \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i)]_{|\phi=\hat{\phi}^I} \right) \\
&\qquad\qquad + \int_0^t Y_i(u) \frac{\alpha_0(u)}{s_0(u)} \mathrm{d}u \\
&\times m^{-1} \sum_{j=1}^{m} \left. \left( \exp\left\{ \beta_0^\top X_{ij}(\hat{\phi}^I) \right\} - E\left[ \exp\left\{ \beta_0^\top X_{i1}(\phi) \right\} \middle| \mathcal{C}_i, G_{\mathcal{C}_i}(Z_i) \right]_{|\phi=\hat{\phi}^I} \right) \right\}
\end{aligned}
\tag{21}
$$

plus $o_P(1)$-terms. Proceeding as in the proof of asymptotic normality of the regression parameters, we can show that the terms in (21) are asymptotically independent of the terms in (20) and converge in distribution to a normal distribution. Also the terms in (20) are asymptotically normal. Thus $n^{1/2}\{\hat{A}(t) - A_0(t)\}$ converges to a Gaussian process with mean 0.

### A.5 Proof of Theorem 1: Joint convergence

To see that $n^{1/2}(\hat{\beta} - \beta_0)$ and $n^{1/2}\{\hat{A}(t) - A_0(t)\}_{t\in[0,\tau]}$ converge jointly in distribution, note that we have written both as a sum of terms—(12), (21)—that depend on the imputations but are asymptotically independent of the observed data, terms—(14), (15), (20)—that depend only on the observed data, and terms, that are asymptotically negligible. Joint convergence follows by noting that linear combinations of the "imputation terms", (12) and (21), are asymptotically independent of the observed data and converge to a normal distribution, while the same linear combinations of the "observed data terms", (14), (15) and (20), also converge to a normal distribution. Hence, $n^{1/2}(\hat{\beta} - \beta_0)$ and $n^{1/2}\{\hat{A}(t) - A_0(t)\}_{t\in[0,\tau]}$ converge jointly in distribution to a Gaussian process.

### A.6 Iterating the estimation process

In order to establish asymptotic results for the iterated multiple-imputation estimator, we extend the arguments in the previous parts of the appendix to the case where the "initial estimator" is a multiple-imputation estimator of the type we are considering. We let $\hat{\phi}^{(1)}$ denote the multiple-imputation estimator based on the initial imputations and let $Z_{ij}^{(2)}(\hat{\phi}^{(1)})$ denote the second iteration imputations, i.e., imputations generated using $\hat{\phi}^{(1)}$ as the true parameter. We focus on the asymptotic distribution of $\hat{\beta}^{(2)}$, the multiple-imputation estimator of $\beta_0$ based on the second iteration imputations and outline the changes we need to make to the expansion of the score function given in Eqs. (12)–(15).

Consider first the term (12). Conditional on the observed data and the first iteration imputations the mean of $S_{\text{eff}}^F\{Z_{ij}^{(2)}(\hat{\phi}^{(1)})\}$ equals $E[S_{\text{eff}}^F\{Z_{ij}^{(2)}(\phi)\}|\mathcal{C}_i, G_{\mathcal{C}}(Z_i)]_{|\phi=\hat{\phi}^{(1)}}$ as the second iteration imputations only depend on the first iteration imputations through the first iteration estimator $\hat{\phi}^{(1)}$. It follows as before that (12) is asymptotically normal and asymptotically independent of the observed data (and the first iteration imputations).

The terms (13) and (14) are unchanged. Finally, the term (15) may be rewritten as $D_{\text{eff}}(\phi_0)n^{1/2}(\hat{\phi}^{(1)} - \phi_0)$. When plugging in the asymptotic expression for $n^{1/2}(\hat{\phi}^{(1)} - \phi_0)$ derived above, and splitting it into the first iteration imputation part corresponding to (12) and (21) and the rest, we end up with a term (12) depending on the second iteration imputations, which is asymptotically independent of the first iteration imputations, terms depending on the first iteration imputations and the observed data, which are asymptotically independent of the observed data, and terms depending only on the observed data. It now follows that the Cox partial score function is asymptotically normal and it is straightforward to verify that it has the same asymptotic distribution as (5) with $q_i$ replaced by $\rho_i = (I^F)^{-1}\xi_i$.

The second iteration estimator of the integrated baseline hazard may be shown to be asymptotically Gaussian by following a similar line of arguments, splitting (21) into a sum of terms depending on the second iteration imputations and terms depending on the first iteration imputations and conditioning as above. Joint convergence follows in a similar manner to what we did for the original multiple-imputation estimator. Further iterations may be handled by splitting the “imputation terms” into additional terms and repeated conditioning.

### A.7 Stochastic equicontinuity

Whereas stochastic equicontinuity of the empirical process based on $m^{-1}\sum_{j=1}^m S_{\text{eff}}^F\{Z_{ij}(\phi_0)\}$ is straightforward to verify when imputing a large class of continuous covariates, we claim that for discrete covariates the combination of the unknown baseline hazard and the inherent discontinuity of the covariate rules out stochastic equicontinuity. To see this, we prove the following lemma:

**Lemma 5** *The set of sets*

$$\left\{ \{(x,t) \in \mathcal{X} \times \mathbb{R} : x \leq a(t)\} \quad a : \mathbb{R} \to \mathbb{R} \text{ increasing} \right\}$$

*with $\mathcal{X} \subset \mathbb{R}$ is a Vapnik–Chervonenkis (VC) class if and only if $\mathcal{X}$ is a finite set.*

***Proof*** Consider a set $A = \{(x_1, t_1), \ldots, (x_n, t_n)\}$. Assuming that $|\mathcal{X}|$ is finite, then any set of $n > |\mathcal{X}|$ points will contain at least two points $(x_i, t_i)$, $(x_j, t_j)$, such that $x_i = x_j$ and (without loss of generality) $t_i \leq t_j$. Clearly, we cannot pick out a subset of $A$ containing $x_i$ but not $x_j$: If $a(t_i) \geq x_i$ then $a(t_j) \geq a(t_i) \geq x_i = x_j$. Thus, no sufficiently large set is shattered, and the set of sets is a VC class. If $\mathcal{X}$ is not finite, then choosing $A$ such that $x_1 < x_2 < \cdots < x_n$ and $t_1 < t_2 < \cdots < t_n$ any subset may be picked out: For a subset $B \subseteq A$ choose $a$ so that it jumps to just above $x_i$ just before $t_i$ for any $i$ such that $(x_i, t_i) \in B$. As $A$ can be shattered, the set of sets is not a VC class. □

Consider imputing a single binary explanatory variable, $X$, with conditional probability of success given by

$$p\{\mathcal{C}, G_{\mathcal{C}}(Z), \phi\} = \frac{\exp\{\Delta\beta - A(T)\exp(\beta)\}p(\theta)}{\exp\{\Delta\beta - A(T)\exp(\beta)\}p(\theta) + \exp\{-A(T)\}\{1 - p(\theta)\}}.$$

Then the simplest way of simulating $X$ is

$$X(\phi) = I[\{\tilde{U} \leq \Delta\beta - A(T)(\exp(\beta) - 1) - \text{logit}\{p(\theta)\}],$$

with $\tilde{U} = \text{logit}(U)$, where $U$ is uniformly distributed. Lemma 5 shows that even if we fix $\beta$ and $\theta$, these indicator functions are not indicators of a VC class of sets. It follows that it is not VC if we allow $\beta$ and $\theta$ to vary, either. Dudley (1984, Theorem 11.4.1) shows that when a set of indicator functions are not based on a VC class, the corresponding empirical process is not pregaussian. This basically rules out stochastic equicontinuity.

This argument shows that the efficient score process with imputed data is not stochastic equicontinuous in general. It does not rule out—though we find it unlikely—that one might construct another simulation scheme which would be sufficiently "smooth" for a discrete covariate to make the process stochastic equicontinuous.

## References


Andersen, P., Borgan, Ø., Gill, R., Keiding, N. (1992). *Statistical models based on counting processes*. New York: Springer.

Bartlett, J., Seaman, S., White, I., Carpenter, J., The Alzheimer's Disease Neuroimaging Initiative. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*, 462–487.

Chen, H. Y. (2002). Double-semiparametric method for missing covariates in Cox regression models. *Journal of the American Statistical Association*, *97*, 565–576.

Chen, H. Y., Little, R. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *94*, 896–908.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statististical Society, Series B*, *34*, 187–220.

Dudley, R. (1984). *A course on empirical processes, volume 1097 of Lecture Notes in Mathematics*. Berlin: Springer.

Herring, A., Ibrahim, J. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Association*, *96*, 292–302.

Jacobsen, M., Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *The Annals of Statistics*, *23*(3), 774–786.

Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. New York: Springer.

Martinussen, T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics*, *26*, 479–491.

Nielsen, S. F. (2003). Proper and improper multiple imputation. *International Statistical Review*, *71*(3), 593–607.

Pugh, M., Robins, J., Lipsitz, S., Harrington, D. (1993). *Inference in the Cox proportional hazards model with missing covariate data*. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Qi, L., Wang, C., Prentice, R. (2005). Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, *100*, 1250–1263.

Robins, J., Wang, N. (2000). Inference for imputation estimators. *Biometrika*, *87*, 113–124.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Schenker, N., Welsh, A. H. (1988). Asymptotic results for multiple imputation. *The Annals of Statistics*, *16*(4), 1550–1566.

Schlichting, P., Christensen, E., Andersen, P., Fauerholdt, L., Juhl, E., Poulsen, H., Tygstrup, N. (1983). Prognostic factors in cirrhosis identified by Cox's regression model. *Hepatology*, *3*, 889–895.

Sterne, J., White, I., Carlin, J., Spratt, M., Royston, P., Kenward, M., Wood, A., Carpenter, J. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, *339*, b2393.

Tsiatis, A. (2006). *Semiparametric theory and missing data*. New York: Springer.

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

van der Vaart, A., Wellner, J. (1996). *Weak convergence and empirical processes. With applications to statistics*. New York: Springer.

Wang, N., Robins, J. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika*, *85*, 935–948.

White, I., Royston, P. (2009). Imputing missing covariate values for the Cox model. *Statistics in Medicine*, *28*, 1982–1998.