Supplementary Materials for “Statistically and computationally efficient conditional mean imputation for censored covariates”

Sarah C. Lotspeich and Ethan M. Alt

S.1 | DERIVATIONS FOR CONDITIONAL MEANS UNDER LEFT AND INTERVAL CENSORING

Suppose that $X > 0$ is a nonnegative random variable with conditional probability density function (PDF) $f(x|\boldsymbol{z})$ and survival function $S(x|\boldsymbol{z})$ given additional fully observed covariates $\boldsymbol{Z} = \boldsymbol{z}$. Suppose that we wish to find the mean of X given $\boldsymbol{Z} = \boldsymbol{z}$ and some bounds on X, say, $0 \leq L < X \leq U < \infty$. Note that this conditional mean is defined:

$$\begin{aligned}
\mathrm{E}(X|L < X \leq U, \boldsymbol{Z}) &= \int_L^U xf(x|\boldsymbol{Z})dx \\
&= \int_L^\infty xf(x|\boldsymbol{Z})dx - \int_U^\infty xf(x|\boldsymbol{Z})dx \\
&= \mathrm{E}(X|X > L, \boldsymbol{Z}) - \mathrm{E}(X|X > U, \boldsymbol{Z}).
\end{aligned} \tag{S.1}$$

Hence, the doubly truncated mean $\mathrm{E}(X|L < X \leq U, \boldsymbol{Z})$ can be expressed as the difference between two singly truncated means with only lower bounds, $\mathrm{E}(X|X > L, \boldsymbol{Z})$ and $\mathrm{E}(X|X > U, \boldsymbol{Z})$.

Define the following conditional PDF for X given covariates $\boldsymbol{Z}$ and a lower bound a on X: $f(x|X > a, \boldsymbol{Z}) = f(x|\boldsymbol{Z})/S(a|\boldsymbol{Z})$ where $x > a$. With this conditional PDF, it follows from the definition of expectation that

$$\mathrm{E}(X|X > a, \boldsymbol{Z}) = \int_a^U xf(x|X > a, \boldsymbol{Z})dx = \frac{\int_a^\infty xf(x|\boldsymbol{Z})dx}{S(a|\boldsymbol{Z})}. \tag{S.2}$$

Then, we can use integration by parts on the numerator in (S.2) to get

$$\int_a^\infty xf(x|\boldsymbol{Z})dx = aS(a|\boldsymbol{Z}) + \int_a^U S(x|\boldsymbol{Z})dx. \tag{S.3}$$

Specifically, to obtain (S.3) we integrate by parts with $u = x$ and $dv = f(x)dx$ so that $du = dx$ and $v = F(x) + b$ for some constant b, where we choose $b = -1$ so that $v = F(x) - 1 = -S(x)$. Plugging (S.3) into (S.2) and simplifying we arrive at

$$\mathrm{E}\left(X|X>a,\boldsymbol{Z}\right)=a+\frac{\int_a^{\infty}S(x|\boldsymbol{Z})dx}{S(a|\boldsymbol{Z})}. \tag{S.4}$$

Notably, (S.4) matches the conditional mean presented for right-censored covariates in (1) of the main text if $a = W_i$.

Finally, combining the expressions in (S.1) and (S.4), we may write the conditional mean for a right-, left-, or interval-censored covariate X in the following piecewise fashion:

$$\mathrm{E}\left(X|L<X\leq U,\boldsymbol{Z}\right)=\begin{cases} L+\frac{\int_L^{\infty}S(x|\boldsymbol{Z})dx}{S(L|\boldsymbol{z})} & \text{if } L\geq 0, U\to\infty \text{ (i.e., } X \text{ is right-censored at } L\text{)},\\ \int_0^{\infty}S(x|\boldsymbol{Z})dx-\left(U+\frac{\int_U^{\infty}S(x|\boldsymbol{Z})dx}{S(U|\boldsymbol{Z})}\right) & \text{if } L=0, U<\infty \text{ (i.e., } X \text{ is left-censored at } U\text{)},\\ \frac{LS(L|\boldsymbol{Z}_i)-US(U|\mathbf{Z})+\int_L^U S(x|\boldsymbol{Z})dx}{S(L|\boldsymbol{Z}_i)-S(U|\mathbf{Z})} & \text{if } 0<L<U<\infty \text{ (i.e., } X \text{ is interval-censored in } (L,U)\text{)}.\end{cases} \tag{S.5}$$

For some distributions and under interval censoring, the following alternative definition may be more convenient when X is interval-censored in (L, U):

$$\mathrm{E}(X|L<X\leq U,\boldsymbol{Z})=\frac{\int_L^U xf(x|\boldsymbol{Z})dx}{S(L|\boldsymbol{Z})-S(U|\boldsymbol{Z})}. \tag{S.6}$$

S.2 | DERIVATIONS FOR POPULAR DISTRIBUTIONS UNDER INTERVAL CENSORING

For all of the derivations that follow, suppose that the covariate X was censored on the interval $(L_i, W_i]$, such that $L_i < X \leq U_i$. All other notation follows from Section 2 of the main text.

S.2.1 | Weibull distribution

Recall from Section 2.4.1 that the nonnegative random variable X has a Weibull distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ if X has PDF $f(x) = \alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha})$ for $x > 0$. The survival function induced by this density is given by $S(x) = \exp(-\lambda x^{\alpha})$ and the hazard function is given by $h(x) = \alpha\lambda x^{\alpha-1}$. Additional fully observed covariates $\boldsymbol{Z}_i$ can be incorporated into the scale parameter λ_i, while the shape parameter α is still assumed to be constant.

Substituting the Weibull survival function into the third case in (S.5) gives the following formula for the conditional mean of an interval-censored covariate:

$$\mathrm{E}\left(X|L_i<X\leq U_i,\boldsymbol{Z}_i\right)=\frac{L_i\exp\left(-\lambda_i L_i^{\alpha}\right)-U_i\exp\left(-\lambda_i U_i^{\alpha}\right)+\int_{L_i}^{U_i}\exp\left(-\lambda_i x^{\alpha}\right)dx}{\exp\left(-\lambda_i L_i^{\alpha}\right)-\exp\left(-\lambda_i U_i^{\alpha}\right)}, \tag{S.7}$$

which is simply the expected value of a Weibull(α, λ_i) random variable truncated from below at L_i and above at U_i. The integral in the numerator of (S.7) is straightforward to compute using numerical integration, particularly when the censoring interval $(L_i, U_i]$ is finite. Still, it can be useful to alternatively express this integral as

$$\int_{L_i}^{W_i} \exp\left(-\lambda_i x^{\alpha}\right) dx = \frac{1}{\alpha\lambda_i^{1/\alpha}} \left(\int_{\lambda_i L_i^{\alpha}}^{\infty} - \int_{\lambda_i U_i^{\alpha}}^{\infty}\right) u^{1/\alpha - 1} \exp(-u) du = \frac{1}{\alpha\lambda_i^{1/\alpha}} \left\{\Gamma(1/\alpha, \lambda_i L_i^{\alpha}) - \Gamma(1/\alpha, \lambda_i U_i^{\alpha})\right\}, \quad \text{(S.8)}$$

where we use the substitution $u = \lambda_i x^{\alpha}$ in the first equality and $\Gamma(a, t) = \int_t^{\infty} x^{a-1} \exp(-x) dx$ is the (upper) incomplete Gamma function in the second. Further, as discussed in the text, the incomplete gamma function in (S.8) can be rewritten as

$$\Gamma(a, t) = \Gamma(a) S_{\Gamma}(t|a), \quad \text{(S.9)}$$

where $S_{\Gamma}(t|a)$ is the survival function of a gamma random variable with shape parameter $a > 0$ and scale parameter equal to 1. Substituting (S.9) into (S.8) and subsequently into (S.7) yields the following computationally convenient, analytically tractable equation for the conditional mean of an interval-censored covariate X assuming a Weibull distribution:

$$\mathrm{E}(X|L_i < X \leq U_i, \boldsymbol{Z}_i) = \frac{L_i \exp\left(-\lambda_i L_i^{\alpha}\right) - U_i \exp\left(-\lambda_i U_i^{\alpha}\right) + \frac{\Gamma(1/\alpha)}{\alpha\lambda_i^{1/\alpha}} \left\{S_{\Gamma}\left(\lambda_i L_i^{\alpha}|1/\alpha\right) - S_{\Gamma}\left(\lambda_i U_i^{\alpha}|1/\alpha\right)\right\}}{\exp\left(-\lambda_i L_i^{\alpha}\right) - \exp\left(-\lambda_i U_i^{\alpha}\right).}. \quad \text{(S.10)}$$

Recognizing that right and left censoring are special cases of interval censoring, we arrive at the analytically tractable solutions for these cases, also. First, letting $L_i = W_i$ and $U_i \to \infty$ in (S.10), we arrive at (5) from the main text, the conditional mean under right censoring. Second, letting $L_i = 0$ and $U_i = W_i$ in (S.10), we derive the following conditional mean under left censoring:

$$\mathrm{E}(X|X < W_i, \boldsymbol{Z}_i) = \frac{\frac{\Gamma(1/\alpha)}{\alpha\lambda_i^{1/\alpha}} \left\{1 - S_{\Gamma}\left(\lambda_i W_i^{\alpha}|1/\alpha\right)\right\} - W_i \exp\left(-\lambda_i W_i^{\alpha}\right)}{1 - \exp\left(-\lambda_i W_i^{\alpha}\right)}.$$

S.2.2 | Log-normal distribution

Recall from Section 2.4.3 that the nonnegative random variable X from a log-normal distribution with location parameter $\mu \in \mathbb{R}$ and scale parameter $\sigma > 0$ has PDF $f(x) = \exp\left[-\{\log(x) - \mu\}^2/\left(2\sigma^2\right)\right]/\left(\sqrt{2\pi}\sigma x\right)$ for $x > 0$. The survival function for a log-normal random variable X is given by $S(x) = \Phi\left[\{\mu - \log(x)\}/\sigma\right]$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Additional covariates $\boldsymbol{Z}_i$ can be incorporated into the location parameter, in which case X given $\boldsymbol{Z}_i$ has a log-normal distribution with location μ_i and scale σ.

By definition, if X has a log-normal distribution (given $\boldsymbol{Z}_i$) with location μ_i and scale σ, then $Y = \log(X)$ has a normal distribution (given $\boldsymbol{Z}_i$) with mean μ_i and variance σ^2. Thus, if we're conditioning on $L_i < X \leq W_i$ from the interval censoring, then we can equivalently condition on $\log(L_i) < Y \leq \log(U_i)$ instead and rewrite $\mathrm{E}\left(X|L_i < X \leq U_i, \boldsymbol{Z}_i\right)$ as

$$\mathrm{E}\left\{\exp(Y)|\log(L_i) < Y \leq \log(U_i), \boldsymbol{Z}_i\right\} = \frac{\int_{\log(L_i)}^{\log(U_i)} \exp(y)\phi(y|\mu_i, \sigma^2) dy}{\Phi\left\{\frac{\log(U_i) - \mu_i}{\sigma}\right\} - \Phi\left\{\frac{\log(L_i) - \mu_i}{\sigma}\right\}}, \quad \text{(S.11)}$$

where $\phi(y|\mu_i, \sigma^2)$ is the PDF of a normal random variable with mean μ_i and variance σ^2 evaluated at $Y = y$. The integral in (S.11) is equal to the moment generating function $M(t) = \mathrm{E}\{\exp(tY)\}$ of the truncated normal distribution for Y (given $\boldsymbol{Z}_i$) evaluated at $t = 1$. Hence, the analytic solution for the conditional mean under interval censoring is

$$\mathrm{E}(X|L_i < X \le U_i, \boldsymbol{Z}_i) = \exp\left(\mu_i + \frac{\sigma^2}{2}\right)\left[\frac{\Phi\left\{\frac{\log(U_i)-\mu_i}{\sigma} - \sigma\right\} - \Phi\left\{\frac{\log(L_i)-\mu_i}{\sigma} - \sigma\right\}}{\Phi\left\{\frac{\log(U_i)-\mu_i}{\sigma}\right\} - \Phi\left\{\frac{\log(L_i)-\mu_i}{\sigma}\right\}}\right], \tag{S.12}$$

which is easy to evaluate using built-in functions for the normal CDF $\Phi(\cdot)$ in standard statistical software (e.g., the `pnorm` function in base-R).

We obtain (8) from the main text, the conditional mean under right censoring, by letting $L_i = W_i$ and $U_i \to \infty$ in (S.12). Additionally, we can derive the conditional mean under left censoring by letting $L_i = 0$ and $U_i = W_i$, so that (S.12) simplifies to the following:

$$\mathrm{E}(X|X < W_i, \boldsymbol{Z}_i) = \exp\left(\mu_i + \frac{\sigma^2}{2}\right)\left[\frac{\Phi\left\{\frac{\log(W_i)-\mu_i}{\sigma} - \sigma\right\}}{\Phi\left\{\frac{\log(W_i)-\mu_i}{\sigma}\right\}}\right].$$

S.2.3 | Log-logistic distribution

Recall from Section 2.4.4. that a nonnegative random variable X from the log-logistic distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$ has its PDF given by $f(x) = (\alpha/\lambda)(x/\lambda)^{\alpha-1}/\{1 + (x/\lambda)^\alpha\}^2$ for $x > 0$, and its survival function defined as $S(x) = 1/\{1 + (x/\lambda)^\alpha\}$. The additional covariates $\boldsymbol{Z}_i$ can be incorporated through the scale parameter λ_i, while the shape parameter α is estimated directly.

Plugging the log-logistic survival function into the third case of (S.5), we begin with:

$$\mathrm{E}(X|L_i < X \le U_i, \boldsymbol{Z}_i) = \frac{L_i/\{1 + (L_i/\lambda_i)^\alpha\} - U_i/\{1 + (U_i/\lambda_i)^\alpha\} + \int_{L_i}^{U_i} 1/\{1 + (x/\lambda_i)^\alpha\}\,dx}{1/\{1 + (L_i/\lambda_i)^\alpha\} - 1/\{1 + (U_i/\lambda_i)^\alpha\}}. \tag{S.13}$$

The key integral in (S.13) can rewritten as:

$$\int_{L_i}^{U_i} \frac{1}{1 + (x/\lambda_i)^\alpha}\,dx = \frac{\lambda_i}{\alpha}\int_{\{1+(U_i/\lambda_i)\}^{-1}}^{\{1+(L_i/\lambda_i)\}^{-1}} u^{-1/\alpha}(1-u)^{1/\alpha-1}\,du, \tag{S.14}$$

where we use the substitution $u = [1 + (x/\lambda_i)^\alpha]^{-1}$ so that $x = \lambda_i(1/u-1)^{1/\alpha}$ and $dx = -(\lambda_i/\alpha)(1-u)^{1/\alpha-1}u^{-(1/\alpha+1)}\,du$. For ease of notation going forward, denote the integral bounds in (S.14) as $U_i^* = \{1 + (U_i/\lambda)^\alpha\}^{-1}$ and $L_i^* = \{1 + (L_i/\lambda)^\alpha\}^{-1}$. We can express (S.14) as the following difference of two integrals:

$$\frac{\lambda_i}{\alpha}\int_{U_i^*}^{L_i^*} u^{-1/\alpha}(1-u)^{1/\alpha-1}\,du = \frac{\lambda_i}{\alpha}\left\{\int_0^{L_i^*} u^{-1/\alpha}(1-u)^{1/\alpha-1}\,du - \int_0^{U_i^*} u^{-1/\alpha}(1-u)^{1/\alpha-1}\,du\right\}. \tag{S.15}$$

Then, using the incomplete gamma function $B(t|a,b) = \int_0^t x^{a-1}(1-x)^{b-1}\,dx$ ($t \in [0,1]$, $a > 0$, $b > 0$) we can rewrite the integrals in (S.15) more concisely as

$$\frac{\lambda_i}{\alpha}\left\{B\left(L_i^*\middle|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right) - B\left(U_i^*\middle|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\right\}.$$

As noted in (10) in the main text, $B(t|a,b) = B(a,b)F_\beta(t|a,b)$, where $F_\beta(t|a,b)$ is the CDF of a beta random variable with shape parameters a and b ($t \in [0,1]$, $a > 0$, $b > 0$) evaluated at t. Finally, using this substitution we have that the integral can be calculated:

$$\int_{L_i}^{U_i} \frac{1}{1+(x/\lambda_i)^\alpha}\,dx = \left(\frac{\lambda_i}{\alpha}\right) B\left(\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\left\{F_\beta\left(L_i^*\middle|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right) - F_\beta\left(U_i^*\middle|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\right\}. \tag{S.16}$$

With (S.16), the analytically tractable equation for the conditional mean of an interval-censored covariate X assuming a log-logistic distribution with shape $\alpha > 0$ is

$$\mathrm{E}(X|L_i < X \le U_i, \boldsymbol{Z}_i) = \frac{L_i L_i^* - U_i U_i^* + \left(\frac{\lambda_i}{\alpha}\right) B\left(\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\left\{F_\beta\left(L_i^*|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right) - F_\beta\left(U_i^*|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\right\}}{L_i^* - U_i^*}. \tag{S.17}$$

For shapes $\alpha \le 1$, the conditional mean $\mathrm{E}(X|L_i < X \le U_i, \boldsymbol{Z}_i)$ does not exist. All of the terms in (S.17) can be efficiently computed using built-in functions of R, such as `beta` and `pbeta`.

The conditional mean under right censoring in (11) of the main text can be obtained by letting $L_i = W_i$, $U_i \to \infty$, and $U_i^* \to 0$ in (S.17). Additionally, we can derive the conditional mean under left censoring by letting $L_i = 0$, $L_i^* = \{1 + (0/\lambda_i)^\alpha\}^{-1} = 1$, and $U_i = W_i$ in (S.17):

$$\mathrm{E}(X|X < W_i, \boldsymbol{Z}_i) = \left(\frac{W_i}{\lambda_i}\right)^\alpha \left\{\left(\frac{\lambda_i}{\alpha}\right)\{1 + (W_i/\lambda_i)^\alpha\} B\left(\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right)\left(1 - F_\beta\left[\left\{1 + \left(\frac{W_i}{\lambda_i}\right)^\alpha\right\}^{-1}\middle|\frac{\alpha-1}{\alpha}, \frac{1}{\alpha}\right]\right) - W_i\right\}.$$

S.2.4 | Piecewise exponential distribution

Suppose that, as in Section 2.4.5, we discretize the axis for a nonnegative random variable X into J disjoint sub-intervals, say, $[\tau_{j-1}, \tau_j)$ for $j \in \{1, \dots, J\}$ and $0 = \tau_0 < \tau_1 < \cdots < \tau_J = \infty$. The interval boundaries τ_j are fixed, and percentiles of the non-censored event times are a common way to choose them. Let λ_j denote the constant hazard corresponding to interval $[\tau_{j-1}, \tau_j)$, i.e., $\lambda_j = h(x)$ for $\tau_{j-1} \le x < \tau_j$. Finally, let $R_j(x)$ denote the time at risk in the j^{th} interval, i.e.,

$$R_j(x) = \begin{cases} 0 & \text{if } x < \tau_{j-1} \\ x - \tau_{j-1} & \text{if } \tau_{j-1} \le x < \tau_j \\ \tau_j - \tau_{j-1} & \text{if } x \ge \tau_j, \end{cases}$$

where $\tau_j - \tau_{j-1}$ is the length of interval j.

Then, a nonnegative random variable X from the piecewise exponential distribution with rate parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_J)$ has PDF

$$f(x) = \prod_{j=1}^{J} \lambda_j^{\mathrm{I}(\tau_{j-1} \le x < \tau_j)} \exp\{-\lambda_j R_j(x)\}$$

for $x > 0$ and $\mathrm{I}(\tau_{j-1} \le x < \tau_j)$ an indicator of x falling inside of sub-interval $[\tau_{j-1}, \tau_j)$. Further, its survival function is given by

$$S(x) = \prod_{j=1}^{J} \exp\left\{-\lambda_j R_j(x)\right\} = \exp\left\{-\sum_{j=1}^{J} \lambda_j R_j(x)\right\}. \tag{S.18}$$

Additional covariates $\boldsymbol{Z}_i$ can be incorporated into modeling the rate parameters by denoting the rate parameters by $\lambda_{ij} = \lambda_j \exp\left(\boldsymbol{Z}_i'\boldsymbol{\beta}\right)$ to indicate the covariates $\boldsymbol{Z}_i$ and the sub-interval j.

We begin by plugging the PWE survival function into the third case of (S.5):

$$\mathrm{E}(X|L_i < X \le U_i, \boldsymbol{Z}_i) = \frac{L_i \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(L_i)\right\} - U_i \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(U_i)\right\} + \int_{L_i}^{U_i} \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(x)\right\} dx}{\exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(L_i)\right\} - \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(U_i)\right\}}. \tag{S.19}$$

Then, we simplify the integral term in (S.19) as follows. We alternatively express the conditional survival function based on (S.18) as

$$S(x|\boldsymbol{Z}_i) = \prod_{j=1}^{J} \left[\exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} \exp\left\{-\lambda_{ij}(x - \tau_{j-1})\right\}\right]^{\mathrm{I}(\tau_{j-1} \le x < \tau_j)}, \tag{S.20}$$

where $H(\tau_j|\boldsymbol{Z}_i) = \sum_{g=1}^{j} \lambda_{ig}(\tau_g - \tau_{g-1})$ is the cumulative hazard function evaluated at sub-interval end point τ_j. Notably, $H(\tau_j|\boldsymbol{Z}_i)$ does not depend on the argument x. For the integral in (S.19), we will use the identity:

$$\int_{L_i}^{U_i} S(x|\boldsymbol{Z}_i)dx = \int_{0}^{U_i} S(x|\boldsymbol{Z}_i)dx - \int_{0}^{L_i} S(x|\boldsymbol{Z}_i)dx. \tag{S.21}$$

The two integrals on the right-hand side of this identity can be simplified more easily using (S.20):

$$\begin{aligned}
\int_0^{U_i} S(x|\boldsymbol{Z}_i) &= \int_0^{U_i} \prod_{j=1}^{J} \left[\exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} \exp\left\{-\lambda_{ij}(x - \tau_{j-1})\right\}\right]^{\mathrm{I}(\tau_{j-1} \le x < \tau_j)} dx, \\
&= \sum_{j=1}^{J_{U_i}} \exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} \int_{\tau_{j-1}}^{\min(U_i,\tau_j)} \exp\left\{-\lambda_{ij}(x - \tau_{j-1})\right\} dx, \\
&= \sum_{j=1}^{J_{U_i}} \exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} \left(\frac{1 - \exp\left[-\lambda_{ij}\left\{\min(U_i,\tau_j) - \tau_{j-1}\right\}\right]}{\lambda_{ij}}\right), \\
&= \sum_{j=1}^{J_{U_i}} \left(\frac{\exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} - \exp\left[-\lambda_{ij}\left\{\min(U_i,\tau_j) - \tau_{j-1}\right\} - H(\tau_{j-1}|\boldsymbol{Z}_i)\right]}{\lambda_{ij}}\right), \\
&= \sum_{j=1}^{J_{U_i}} \left(\frac{\exp\left\{-H(\tau_{j-1}|\boldsymbol{Z}_i)\right\} - \exp\left[-H\left\{\min(U_i,\tau_j)\right\}|\boldsymbol{Z}_i\right]}{\lambda_{ij}}\right), \\
&= \sum_{j=1}^{J_{U_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(U_i,\tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}}\right]
\end{aligned} \tag{S.22}$$

where J_{U_i} denotes the interval into which U_i falls, i.e., $J_{U_i} = \{j : \tau_{j-1} \le U_i < \tau_j\}$. The same steps can be used to obtain the other integral:

$$\int_0^{L_i} S(x|\boldsymbol{Z}_i) = \sum_{j=1}^{J_{L_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(L_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right], \tag{S.23}$$

where J_{L_i} denotes the interval into which L_i falls, i.e., $J_{L_i} = \{j : \tau_{j-1} \leq L_i < \tau_j\}$.

Plugging (S.22) and (S.23) into the identity in (S.21) we have

$$\int_{L_i}^{U_i} S(x|\boldsymbol{Z}_i)dx = \sum_{j=1}^{J_{U_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(U_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right] - \sum_{j=1}^{J_{L_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(L_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right], \tag{S.24}$$

which can be used in (S.19) to construct the analytical solution for an interval-censored X assuming a piecewise exponential distribution:

$$\begin{aligned}\mathrm{E}(X|L_i < X \leq U_i, \boldsymbol{Z}_i) = &\frac{L_i \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(L_i)\right\} - U_i \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(U_i)\right\}}{\exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(L_i)\right\} - \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(U_i)\right\}} \\ &+ \frac{\sum_{j=1}^{J_{U_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(U_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right] - \sum_{j=1}^{J_{L_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(L_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right]}{\exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(L_i)\right\} - \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(U_i)\right\}}.\end{aligned} \tag{S.25}$$

The rate parameters $\boldsymbol{\lambda}$ for (S.25) can be estimated using the `pchreg` function in the *eha* package [1].

The conditional mean under right censoring in (14) of the main text can be obtained by letting $L_i = W_i$ and $U_i \to \infty$ in (S.25). Additionally, we can derive the conditional mean under left censoring by letting $L_i = 0$ and $U_i = W_i$ such that (S.25) simplifies to:

$$\mathrm{E}(X|X < W_i, \boldsymbol{Z}_i) = \frac{\sum_{j=1}^{J_{W_i}} \left[\frac{S(\tau_{j-1}|\boldsymbol{Z}_i) - S\left\{\min(W_i, \tau_j)|\boldsymbol{Z}_i\right\}}{\lambda_{ij}} \right] - W_i \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(W_i)\right\}}{1 - \exp\left\{-\sum_{j=1}^{J} \lambda_{ij} R_j(W_i)\right\}}.$$

references

[1] Broström G. Event History Analysis with R, Second Edition. Boca Raton: Chapman and Hall/CRC; 2022.

S.3 | ADDITIONAL FIGURES

FIGURE S1 Total computing runtime across 1000 replicates for single imputation simulations (in seconds)

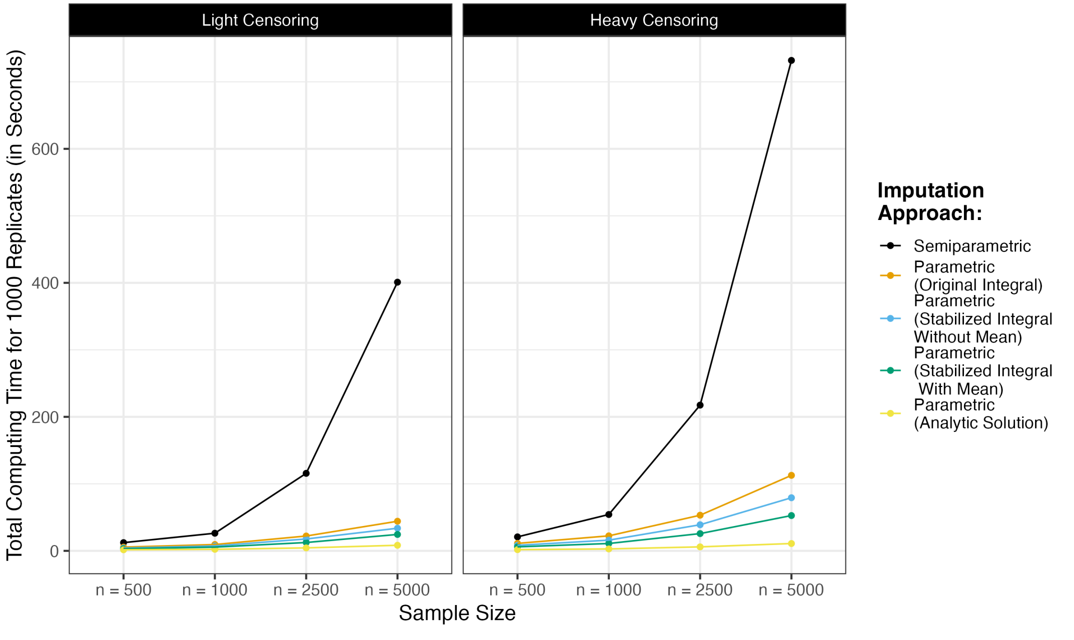

FIGURE S2 Estimates of β_1, the parameter on the censored covariate X in the linear regression analysis model, resulting from each single imputation approach. The horizontal dashed line denotes the true value of $\beta_1 = 0.5$.

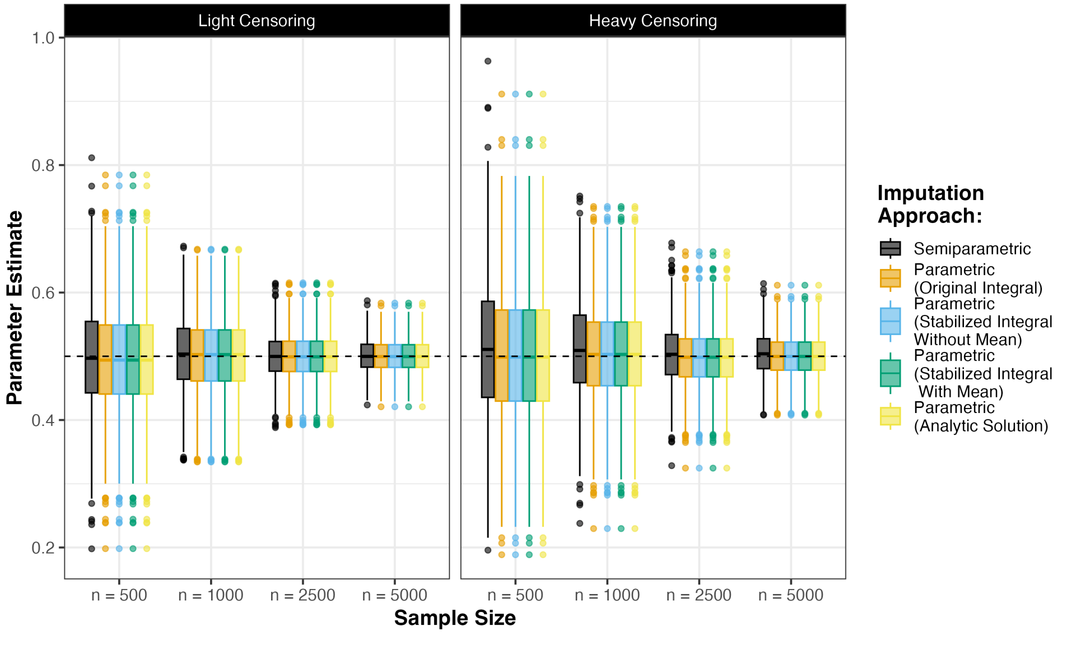

FIGURE S3 Total computing runtime across 1000 replication for imputation simulations (in seconds) with an increasing number of imputations B. Censoring was heavy, and $n = 1000$ subjects were simulated per replicate.

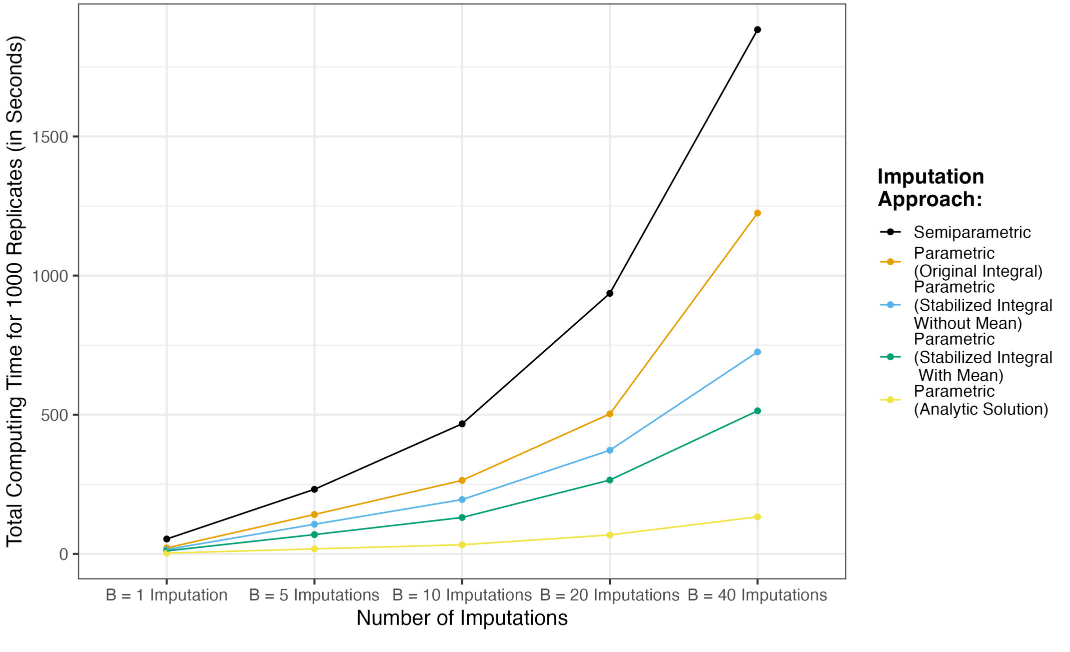

FIGURE S4 Estimates of β_1, the parameter on the censored covariate X in the linear regression analysis model, resulting from each imputation approach with an increasing number of imputations B. The horizontal dashed line denotes the true value of $\beta_1 = 0.5$. Censoring was heavy, and $n = 1000$ subjects were simulated per replicate.

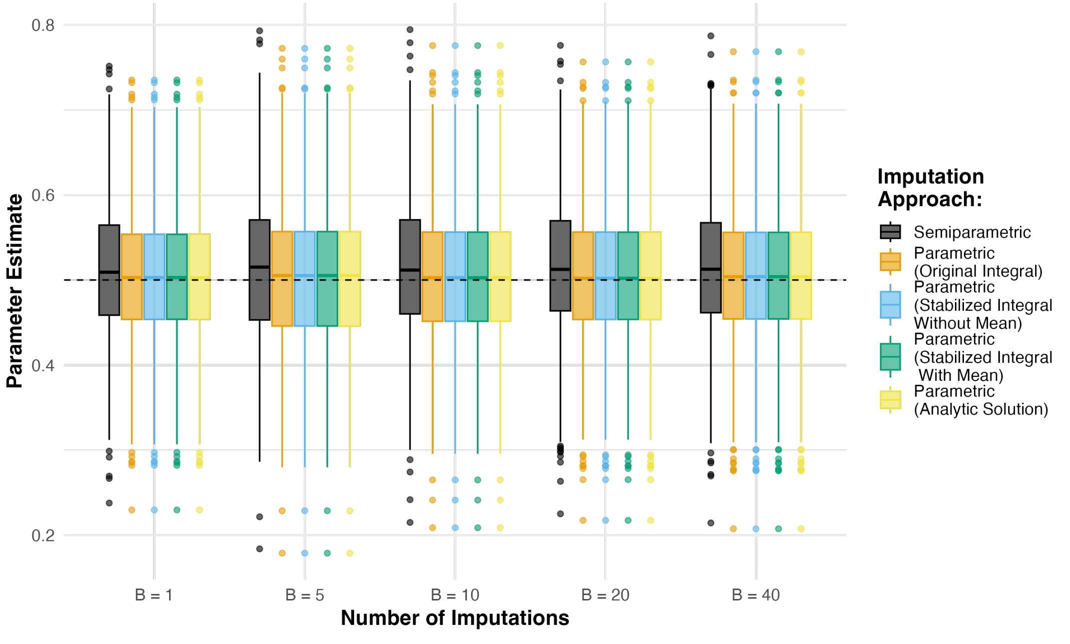

FIGURE S5 Empirical density of time from first visit to hypertension diagnosis (observed or singly imputed) in the Framingham teaching dataset using various distributions for the imputation model. The vertical dashed line denotes $TIME = 25$ years to hypertension diagnosis, which was the end of follow-up.