Paul W. Bernhardt[1]

Maximum Likelihood Estimation in a Semicontinuous Survival Model with Covariates Subject to Detection Limits

[1] Villanova University, Villanova, PA, USA, E-mail: paul.bernhardt@villanova.edu

Abstract:
Semicontinuous data are common in biological studies, occurring when a variable is continuous over a region but has a point mass at one or more points. In the motivating Genetic and Inflammatory Markers of Sepsis (GenIMS) study, it was of interest to determine how several biomarkers subject to detection limits were related to survival for patients entering the hospital with community acquired pneumonia. While survival times were recorded for all individuals in the study, the primary endpoint of interest was the binary event of 90-day survival, and no patients were lost to follow-up prior to 90 days. In order to use all of the available survival information, we propose a two-part regression model where the probability of surviving to 90 days is modeled using logistic regression and the survival distribution for those experiencing the event prior to this time is modeled with a truncated accelerated failure time model. We assume a series of mixture of normal regression models to model the joint distribution of the censored biomarkers. To estimate the parameters in this model, we suggest a Monte Carlo EM algorithm where multiple imputations are generated for the censored covariates in order to estimate the expectation in the E-step and then weighted maximization is applied to the observed and imputed data in the M-step. We conduct simulations to assess the proposed model and maximization method, and we analyze the GenIMS data set.

Keywords: cure model, detection limit, mixture of normals, Monte Carlo EM algorithm, semicontinuous data, survival analysis

DOI: 10.1515/ijb-2017-0058

Received: July 26, 2017; **Revised:** June 22, 2018; **Accepted:** September 28, 2018

1 Introduction

Semicontinuous data are common in medical and biological studies and typically occur when there is a point mass at zero and a continuous, right-skewed distribution over a positive range. Researchers have analyzed this type of data in a variety of applications, including annual medical expenses [1, 2], daily alcohol intake [3], driving scores [4], and health assessment scores [5]. Manning et al. [6], Duan et al. [7], Moulton et al. [8] and Olsen and Schafer [9], among many others, have suggested two-part mixture models for handling both cross-sectional and longitudinal semicontinuous data. In the presence of covariates, the two-part model usually employs logistic regression to model the probability of observing a zero outcome and a continuous regression model to describe the distribution of outcomes above zero.

In the motivating Genetic and Inflammatory Markers of Sepsis (GenIMS) Study, biological measurements were collected on patients admitted to the hospital with community acquired pneumonia (CAP). One of the main purposes of the study was to determine how three cytokine biomarkers are related to survival for patients with CAP. Unlike typical analyses of time-to-event data, the original study investigators were primarily interested in modeling the binary event of surviving at least 90 days [10], based on recommendations by two international expert panels, which suggested that most if not all deaths due to pneumonia and resulting sepsis would have occurred by this time [11, 12]. Alternatively, some recent analyses of the data have directly modeled the available survival times [e.g., [13–15].

Rather than choosing between a binary and continuous survival outcome, we suggest simultaneously modeling the event of surviving 90 days and the actual survival times for those not surviving to 90 days. In order to frame this problem in a familiar way, we note that for the purposes of the GenIMS study, we may conceptually consider the patients who survived to 90 days as cured from CAP. Though cure rate models have been developed to model long-term survival data [16, 17], a typical cure rate model does not apply for the GenIMS

Paul W. Bernhardt is the corresponding author.

data set since the cure status is known for all individuals and the survival part of the model is truncated. Thus, we instead suggest framing the problem in the context of a semicontinuous regression model, which in this survival modeling context we term a semicontinuous survival model. Unlike in a typical semicontinuous data model, the continuous part of the distribution occurs between zero and 90, while the point mass occurs at 90.

An additional complication with the GenIMS data set is that the biomarker covariates of interest are subject to lower detection limits (DLs). Traditionally, a few common approaches have been taken to handle censoring due to DLs. One approach has been to replace the censored observations using either a function of the DL, such as $\mathrm{DL}/\sqrt{2}$ [see [18], or the conditional mean of the censored covariate [19, 20]. These naive substitution methods have been shown to lead to biased parameter estimates in generalized linear models [21–27] and survival models [13–15]. An alternative approach only uses complete cases. While this strategy still leads to unbiased parameter estimates as long as censoring is only caused by the DL, efficiency is lost due to the discarding of data.

Several authors have considered more sophisticated approaches to handling DLs in the context of survival data. Langohr et al. [28] and Sattar et al. [15] proposed fully-parametric survival models with an interval-censored predictor while Bernhardt et al. [13] used multiple imputation techniques for accelerated failure time models with left-censored covariates. In the context of semiparametric Cox proportional hazards models with covariates subject to DLs, Lee et al. [29] suggested a partial likelihood approach based on an empirical estimate of the relative risk function using uncensored covariate observations, D' Angelo and Weissfeld [14] developed an indexing approach where censored covariate values are directly replaced by their conditional expectation given a linear combination of the fully observed covariates, and Chen et al. [30] considered a Bayesian strategy where the censored covariates are modeled parametrically. Recently, [31] suggested using a nonparametric density estimator for a single censored covariate in the context of frailty models.

All of the methods mentioned above for handling covariates subject to DLs in survival models are either statistically naive — such as the $\mathrm{DL}/\sqrt{2}$ and complete case methods — or do not consider a model set-up where individuals are no longer considered to experience the event of interest after a fixed time point. In the context of semicontinuous models, which we propose to model this special type of data, the issue of covariates subject to DLs has not been previously considered. Additionally, to our knowledge, prior statistical analyses using semicontinuous regression models have not included a case where the continuous portion of the distribution is bounded above.

In this paper, we propose a straightforward and flexible method for analyzing a semicontinuous survival outcome with covariates subject to DLs, where the interest lies in estimating both the associations of covariates with the survival time for those experiencing the event of interest and the associations of covariates with the probability of surviving to the known "cure time." We suggest using a truncated accelerated failure time model to model the survival times for those experiencing the event of interest and a logistic regression to model the probability of being cured (surviving to the "cure time"). We propose using a series of mixture of normal regression models to flexibly model the multivariate distribution of the censored covariates, though our modeling strategy allows for a more standard parametric model or even a semiparametric approach. To obtain estimates in this semicontinuous survival model, we develop a Monte Carlo EM algorithm where to estimate the expectation in the E-step, multiple imputations are generated for the censored covariates according to a mixture of normals distribution, and weighted maximum likelihood is applied to the observed and imputed data.

The remainder of the paper is organized as follows. In Section 2, we develop our proposed semicontinuous survival model, including our proposal for the distribution of the covariates. In Section 3, we present a Monte Carlo EM algorithm for obtaining parameter estimates. In Section 4, we summarize the main goals of inference as well as some notes on model checking. In Section 5, we conduct a simulation study comparing our proposed method to naive or less flexible methods. In Section 6, we apply our method to the GenIMS data set. Finally, in Section 7, we briefly review our method and discuss its advantages and shortcomings.

2 Semicontinuous survival model

2.1 Notation

Suppose for each individual, $i = 1, \cdots, n$, we observe $\{Y_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i, \boldsymbol{\delta}_i\}$, where Y_i is a time-to-event variable that is continuous over the range $[0,c)$ and has a probability mass at c, $\boldsymbol{X}_i^* = (X_{i1}^*, X_{i2}^*, \cdots, X_{iq}^*)^T = (\max\{X_{i1}, d_{i1}\}, \max\{X_{i2}, d_{i2}\}, \cdots, \max\{X_{iq}, d_{iq}\})^T$, where $\boldsymbol{X}_i$ is a $q \times 1$ vector of covariates subject to the lower DLs $\boldsymbol{d}_i = (d_{i1}, d_{i2}, \cdots, d_{iq})^T$, $\boldsymbol{Z_i}$ is a $(p-q) \times 1$ vector of fully observed covariates, and $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2}, \cdots, \delta_{iq})^T$ is a vector

of censoring indicators such that

$$\delta_{ij} = \begin{cases} 1 & \text{if } X_{ij} \geq d_{ij} \\ 0 & \text{if } X_{ij} < d_{ij} \end{cases} \quad j = 1, \cdots, q.$$

For the ith individual, we additionally define $\boldsymbol{X}_i^{obs}$ as the subset of $\boldsymbol{X}_i$ for which $\delta_{ij} = 1$ and $\boldsymbol{X}_i^{cen}$ and $\boldsymbol{d}_i^{cen}$ as the subsets of $\boldsymbol{X}_i$ and $\boldsymbol{d}_i$, respectively, for which $\delta_{ij} = 0$.

We allow the vector of DLs $\boldsymbol{d}_i$ to be individual specific for generality purposes and also because one of the biomarkers of interest in the motivating GenIMS study is measured with two different values for the DL. Also, while we focus on covariates subject to lower DLs, as motivated by the GenIMS study, the method we suggest can be easily generalized to handle covariates censored due to upper DLs, or a combination of both types. Finally, while in this paper we consider the point mass to exist at c since in the GenIMS analysis patients are effectively considered cured at time c, the model we present in the next section could be used to describe a variable with a point mass at any point $\geq c$. For example, in cure modeling literature, it is typical to consider the point mass as philosophically occurring at infinity since the survival time for cured individuals can never be observed.

2.2 Model

The main inferential goals for a semicontinuous survival outcome are three-fold: (a) to determine the relationship between the covariates and the probability of surviving to time c; (b) to determine the relationship between the covariates and the survival time for those individuals not surviving to time c; (c) to determine if there is a statistically significant relationship between the covariates and the semicontinuous survival outcome. In order to address these goals, we first define the semicontinuous survival density $f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$ with mass at c as follows:

$$f(y_i|\boldsymbol{x}_i, z_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = \{p(y_i|\boldsymbol{x}_i, z_i; \boldsymbol{\gamma}) g(y_i|\boldsymbol{x}_i, z_i; \boldsymbol{\beta}, \boldsymbol{\eta})\}^{\mathbb{1}(y_i<c)} \{1 - p(y_i|\boldsymbol{x}_i, z_i; \boldsymbol{\gamma})\}^{\mathbb{1}(y_i=c)}, \tag{1}$$

where $p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma})$ is modeled via logistic regression so that $P(Y_i < c) = p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma}) = [1 + \exp\{-(1, \boldsymbol{x}_i^T, \boldsymbol{z}_i^T)\boldsymbol{\gamma})\}]^{-1}$, $g(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta})$ is a truncated distribution with support $[0,c)$, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the $(p+1)\times 1$ parameter vectors of primary interest. While the survival time Y_i is not subject to censoring in the GenIMS study, to keep within the context of standard survival methods, it may be convenient to model Y_i using an accelerated failure time (AFT) model approach, so that

$$\log(Y_i) = (1, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)\boldsymbol{\beta} + \sigma\epsilon_i,$$

where ϵ_i is distributed so that $Y_i|\boldsymbol{X}_i, \boldsymbol{Z}_i \sim g(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta})$, and $\boldsymbol{\eta}$ is a parameter vector including both σ and any other parameters in the distribution of ϵ_i. Unlike a standard AFT model, $g(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$ in eq. (1) has a truncated support. However, we still suggest adopting a typical flexible AFT model distribution such as the Weibull or generalized gamma. We note that the sets of covariates included in the logistic and AFT parts of the model do not necessarily need to be the same, though any covariate for which the overall survival relationship is of interest should be included in both models. Since having separate sets of covariates for the logistic and AFT models is a fairly trivial modeling adjustment and mainly complicates notation, we do not address this scenario henceforth.

The observed data likelihood for the proposed semicontinuous model can be written succinctly as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\boldsymbol{d}_i^{cen}} f(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) f(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}|\boldsymbol{z}_i; \boldsymbol{\theta}_{\boldsymbol{x}}) d\boldsymbol{x}_i^{cen}, \tag{2}$$

where $f(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\theta}_{\boldsymbol{x}})$ is the conditional distribution of the covariates $\boldsymbol{X}_i$ subject to censoring with parameter vector $\boldsymbol{\theta}_{\boldsymbol{x}}$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T, \boldsymbol{\gamma}^T, \boldsymbol{\theta}_{\boldsymbol{x}}^T)^T$ is the vector of all of the parameters in the model. The dimension of the integral in eq. (2) is individual specific and corresponds to the number of covariates in $\boldsymbol{X}_i$ observed to be censored below the DLs $\boldsymbol{d}_i$ (the length of the $\boldsymbol{X}_i^{cen}$ vector). For those individuals where all q of the covariates subject to DLs are censored below $\boldsymbol{d}_i$, the dimension of the integral is q. On the other hand, for those individuals without any censored covariates, the contribution to the likelihood eq. (2) simplifies to $f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) f(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\theta}_{\boldsymbol{x}})$. We briefly note that it is not necessary to include a missingness model in the likelihood eq. (2) since, given the observed vector of censoring indicators $\boldsymbol{\delta}_i$, we know with certainty whether or not a value is missing.

For a univariate X_i that is not known to follow a standard distribution, we suggest modeling $f(x_i|\boldsymbol{z}_i)$ using a mixture of normal regressions,

$$f(x_i|\boldsymbol{z}_i; \boldsymbol{\theta}_x) = \sum_{k=1}^{s} \pi_k \phi\{x_i; \mu(\boldsymbol{z}_i, \boldsymbol{\alpha}_{\boldsymbol{k}}), \tau_k\}, \tag{3}$$

where $\boldsymbol{\theta}_x = (\boldsymbol{\pi}^T, \boldsymbol{\alpha}^T, \boldsymbol{\tau}^T)^T$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_s)^T$ is a vector of mixture probabilities, $\phi\{x_i; \mu(\boldsymbol{z}_i, \boldsymbol{\alpha_k}), \tau_k\}$ is a normal distribution with mean $\mu_{ik} = (1, \boldsymbol{z}_i^T)\boldsymbol{\alpha_k}$ and variance τ_k, $k = 1, \cdots, s$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha_1}^T, \boldsymbol{\alpha_2}^T, \cdots, \boldsymbol{\alpha_s}^T)^T$ is an $\{s \cdot (p - q + 1)\} \times 1$ vector of coefficients for the regression of X_i on $\boldsymbol{Z}_i$.

Mixtures of normal regression models were introduced by Quandt and Ramsey [32] and have been shown to be suitably flexible to represent a wide variety of densities [33]. If desired, to increase flexibility of the mixture distribution (3), $\pi_1, \cdots, \pi_s$ can be modeled using a multinomial logistic regression on $\boldsymbol{Z}_i$, though we do not take this approach here. In practice, due to the censoring of the left tail of the distribution for X_i, we suggest letting s be relatively small, especially if the censoring percentage is high. Even with $s = 2$, mixtures of normals can well represent a wide variety of densities. However, as with any parametric distributional assumption, care must be taken. When the covariate censoring rate is high or a mixture of normals is inappropriate for describing the data, identifiability issues may potentially arise, and a simpler model for $f(x_i|\boldsymbol{z}_i)$ may be preferable. We note, however, that identifiability is less of a concern for the submodels of the semicontinuous response. The Monte Carlo EM algorithm that we propose in the next section relies on obtaining imputations for the censored covariates in order to approximate integration in the the E-step, but once imputations are obtained, the binary and continuous components of the semicontinuous regression model can be maximized on the basis of complete data. Of course, perfect separation is always a possibility with binary regression models.

For a multivariate $\boldsymbol{X}_i$, rather than modeling the multivariate distribution of $\boldsymbol{X}_i$ directly, we suggest using the strategy of Ibrahim [34] to model a series of univariate distributions, each using a mixture of normal regressions as in eq. (3), by noting that

$$f(\boldsymbol{x}_i|\boldsymbol{z}_i) = f(x_{iq}|x_{i1}, \cdots, x_{i,q-1}, \boldsymbol{z}_i) f(x_{i,q-1}|x_1, \cdots, x_{i,q-2}, \boldsymbol{z}_i) \cdots f(x_{i1}|\boldsymbol{z}_i). \tag{4}$$

3 Monte carlo EM algorithm

Unless q is small, eq. (2) may be extremely challenging to maximize directly using algorithms like the Newton-Raphson method, since a potentially high-dimensional integral would need to be evaluated for each individual at each iteration in the maximization procedure. Thus, in order to maximize the likelihood eq. (2), we suggest using an expectation-maximization (EM) algorithm, originally described by [35]. The main advantages of using the EM algorithm for maximization over competing methods are that it is very numerically stable [36] and is convenient in the presence of missing data. Maximization is achieved by iterating between two steps, the E-step and M-step. In the E-step, the expected value of the log-likelihood of the complete data is calculated with respect to the missing data conditional on all of the observed data at a set of current parameter estimates. In the M-step, this expected log-likelihood is maximized to obtain new parameter estimates. In this section, we extend the EM algorithm to handle covariates subject to DLs in the context of a semicontinuous survival model.

Suppose that, without loss of generality, we observe $\boldsymbol{\delta}_i = \mathbf{1}$ for the first r individuals and $\delta_{ij} = 0$ for at least one covariate X_{ij} for the remaining $n-r$ individuals. In the E-step of the EM algorithm, the conditional expected value of the log-likelihood with respect to the missing data can be represented by

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \sum_{i=1}^{n} E\left[\log\{f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma})\} + \log\{f(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau})\} \mid y_i, \boldsymbol{x}_i^*, \boldsymbol{z}_i, \boldsymbol{\delta}_i, \boldsymbol{\theta}^{(t)}\right] \\
&= \sum_{i=1}^{r} \left[\log\{f(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma})\} + \log\{f(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau})\}\right] \\
&\quad + \sum_{i=r+1}^{n} \int_{-\infty}^{\boldsymbol{d}_i^{cen}} \big[\log\{f(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}, \boldsymbol{z}_i; \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\gamma})\} \\
&\quad + \log\{f(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}|\boldsymbol{z}_i; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau})\}\big] f(\boldsymbol{x}_i^{cen}|y_i, \boldsymbol{x}_i^{obs}, \boldsymbol{z}_i; \boldsymbol{\theta}^{(t)}) d\boldsymbol{x}_i^{cen} \\
&= \sum_{i=1}^{r} \log\{L(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\} + \sum_{i=r+1}^{n} Q_i^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\}
\end{aligned}$$

A closed form expression of eq. (5) is usually not attainable, so we propose a Monte Carlo approach for estimating this integral, originally suggested by Wei and Tanner [37], in the same spirit as Ibrahim [34] for missing data in parametric regression models and May et al. [38] for generalized linear models with covariates subject to censoring. Specifically, we estimate $Q_i^*(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ as

$$\frac{1}{m} \sum_{l=1}^{m} \log\{L(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i^{obs}, \boldsymbol{x}_{il}, \boldsymbol{z}_i)\}, \tag{6}$$

where $\boldsymbol{x}_{il}, l = 1, \cdots, m$, are randomly sampled from $f_i^T(\boldsymbol{x}_i^{cen}|y_i, \boldsymbol{x}_i^{obs}, \boldsymbol{z}_i; \boldsymbol{\theta}^{(t)})$, the distribution of $\boldsymbol{X}_i^{cen}|Y_i, \boldsymbol{X}_i^{obs}, \boldsymbol{Z}_i$ truncated above at $\boldsymbol{d}_i$. Standard Monte Carlo sampling techniques can be used to obtain samples from this distribution by noting that

$$f(\boldsymbol{x}_i^{cen}|y_i, \boldsymbol{x}_i^{obs}, \boldsymbol{z}_i) \propto f(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}, y_i|\boldsymbol{z}_i) = f(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}, \boldsymbol{z}_i) f(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_i^{cen}|\boldsymbol{z}_i).$$

In practice, we use a random walk Metropolis method [39] with truncated Gaussian proposals.

For the M-step of the EM algorithm, we then maximize eq. (5) with the Monte Carlo estimate eq. (6) replacing $Q_i^*, i = r+1, \cdots, n$. This maximization can be done in three separate parts: a maximization for the part involving the parameters in the logistic regression,

$$\begin{aligned} &\textstyle\sum_{i=1}^{r} \left(\mathbb{1}(y_i < c) \log\{p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma})\} + \mathbb{1}(y_i = c) \log\left[\{1 - p(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\gamma})\}\right] \right) \\ &+ \textstyle\sum_{i=r+1}^{n} \frac{1}{m} \sum_{l=1}^{m} \left(\mathbb{1}(y_i < c) \log\left\{p(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_{il}, \boldsymbol{z}_i; \boldsymbol{\gamma})\right\} + \mathbb{1}(y_i = c) \log\left[\{1 - p(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_{il}, \boldsymbol{z}_i; \boldsymbol{\gamma})\}\right] \right), \end{aligned} \tag{7}$$

a maximization for the part involving the parameters in the distribution of the time-to-event variable for those experiencing an event,

$$\sum_{i=1}^{r} \left[\mathbb{1}(y_i < c) \log\{g(y_i|\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\beta}, \eta)\} \right] + \sum_{i=r+1}^{n} \frac{1}{m} \sum_{l=1}^{m} \left[\mathbb{1}(y_i < c) \log\{g(y_i|\boldsymbol{x}_i^{obs}, \boldsymbol{x}_{il}, \boldsymbol{z}_i; \boldsymbol{\beta}, \eta)\} \right], \tag{8}$$

and a maximization for the part involving the parameters in the distribution of the covariates subject to DLs,

$$\sum_{i=1}^{r} \left[\log\{f(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau})\} \right] + \sum_{i=r+1}^{n} \frac{1}{m} \sum_{l=1}^{m} \left[\log\{f(\boldsymbol{x}_i^{obs}, \boldsymbol{x}_{il}|\boldsymbol{z}_i; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau})\} \right]. \tag{9}$$

The eqs. (7) and (8) can be maximized using standard optimization routines for logistic regression and AFT models, respectively, for a set of $r + (n-r)m$ complete observations, where each of the $(n-r)m$ parts involving the imputed $\boldsymbol{x}_{il}$ values is given weight $1/m$. If a series of mixture of normal regressions is used to model the distribution $f(\boldsymbol{x}_i|\boldsymbol{z}_i)$, we can similarly conduct a series of weighted maximizations for eq. (9) using an EM algorithm for mixture distributions (e.g., [40]). The E and M steps are then iterated until the parameter estimates converge.

We summarize the steps for the proposed Monte Carlo EM algorithm, with additional details and computational tips based on the statistical software `R`.

Step 1: Obtain initial values for estimating the parameters $\{\boldsymbol{\beta}, \eta, \boldsymbol{\gamma}\}$, $\{\hat{\boldsymbol{\beta}}^{(1)}, \hat{\eta}^{(1)}, \hat{\boldsymbol{\gamma}}^{(1)}\}$, by maximizing eq. (2) for the complete cases. The "optim' function can be used to obtain estimates for $\boldsymbol{\beta}$ and η in the truncated AFT model while the function "glm" can be used to obtain estimates for $\boldsymbol{\gamma}$ in the logistic regression model. Obtain initial estimates for the parameters $\{\boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\tau}\}$, $\{\widehat{\boldsymbol{\pi}}^{(1)}, \hat{\boldsymbol{\alpha}}^{(1)}, \widehat{\boldsymbol{\tau}}^{(1)}\}$, by maximizing eq. (3) with censored X_{ij} values replaced by random draws from a normal distribution truncated above at d_{ij}. The function "flexmix" in the `flexmix` package [41] can be used to obtain these estimates.

Step 2: At the $(t+1)$th step of the EM algorithm, given the current set of parameter estimates $\hat{\boldsymbol{\theta}}^{(t)}$, generate m imputed vectors for each $\boldsymbol{X}_i^{cen}$. The function "metrop" in the `mcmc` package [42] can be used to efficiently obtain samples.

Step 3: Update $\hat{\boldsymbol{\theta}}^{(t)}$ as $\hat{\boldsymbol{\theta}}^{(t+1)}$ by maximizing eqs. (7), (8), and (9). The functions flexsurvreg, glm, and flexmix can all handle weighted observations for obtaining maximum likelihood estimates.

Step 4: Repeat steps 2-3 until

$$\max\left\{ \frac{|\hat{\theta}_1^{(t+1)} - \hat{\theta}_1^{(t)}|}{|\hat{\theta}_1^{(t)}|}, \cdots, \frac{|\hat{\theta}_s^{(t+1)} - \hat{\theta}_s^{(t)}|}{|\hat{\theta}_s^{(t)}|} \right\} < \epsilon,$$

where s is the number of elements in $\boldsymbol{\theta}$ and ε is a predefined distance. Define the maximum likelihood parameter estimate as $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(t+1)}$.

In practice, we recommend letting m, the number of Monte Carlo imputations for $\boldsymbol{X}_i^{cen}$, depend on the iteration t by starting with a relatively small m, say 5-15, and increasing m gradually. To obtain variance estimates for $\hat{\boldsymbol{\theta}}$, the method of Louis [43] can be used to estimate the observed information matrix. Alternatively, if it is computationally feasible, standard software can be used to calculate the Hessian matrix for the observed data likelihood eq. (2) at the final set of parameter estimates or bootstrapping can be conducted on the original data set. Finally, we note that the imputations for $\boldsymbol{X}_i^{cen}$ could be obtained in semiparametric or nonparametric manner rather than using a mixture of normal regressions. However, we feel that a parametric approach often takes the most advantage of the data information above the DL in order to extrapolate imputations below the DL.

4 Inference and model checking

Recall that the main analysis goals in the semicontinuous survival regression model are to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the parameters relating the covariates to the continuous and discrete components of the response, and to determine significance of the associations between the covariates and the semicontinuous survival variable.

While estimating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as described in Section 3 is relatively complicated, interpreting $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ is straightforward. Since the binary and continuous parts of the survival model do not share parameters, they can be considered separately (simultaneous maximization is only required due to the censoring on the covariates). Then, in the logistic regression submodel, $\exp\{\gamma_j\}, j = 1, \cdots, p+1$, represent the multiplicative changes in odds of surviving to time c, while in the AFT submodel, $\exp\{\beta_j\}, j = 1, \cdots, p+1$, represent the multiplicative changes in survival times for those experiencing the event of interest before time c.

To assess whether the jth covariate is associated with the survival outcome, the hypothesis test of interest is $H_0 : \beta_j = \gamma_j = 0$ versus $H_a : \beta_j \neq 0$ or $\gamma_j \neq 0$. Since the proposed model is entirely parametric, it is reasonable to calculate the likelihood ratio statistic $-2\log\{L_{H_0}(\hat{\theta})/L_{H_a}(\hat{\theta})\}$ and compare it to the $(1-\alpha)$th quantile of a chi-squared distribution with 2 degrees of freedom.

With the proposed parametric modeling scheme, it is important to choose a good model prior to conducting inference. In practice, we suggest fitting several competing models and comparing them using standard tools for assessing fit such as AIC or BIC. For modeling the binary event of surviving to time c, two popular choices are probit and logistic regression, while for modeling the observed survival times less than c, lognormal, exponential, Weibull, gamma, and generalized gamma models, among others, could be considered. Even if these response models are reasonable, a poor imputation model for the censored covariates could bias inference. While we suggested a flexible mixture of normal regressions approach, alternative distributions could be considered. However, we show in Section 5 that the mixture of normals performs well in a variety of scenarios.

After choosing the best model among those considered, residual analyses or goodness-of-fit tests may be of interest to further confirm that the best model is in fact a reasonable model. However, diagnostics are complicated by the censoring on $\boldsymbol{X}$. One obvious choice for constructing residuals or fitted responses in this scenario is to simply replace the censored values of a covariate by the average of the m values generated at the last iteration of the Monte Carlo EM algorithm. Alternatively, as suggested by [44] and [45], these m imputations could be used to form m multiply imputed data sets after which usual model validation techniques can be conducted repeatedly and summarized appropriately.

5 Simulation study

We conducted simulations to study the performance of our proposed semicontinuous survival model. We set up the simulations to represent a situation similar to that in the GenIMS application described in Section 6, and we compared the proposed method to other practical approaches.

5.1 Single censored covariate: response model known

We generated $N = 500$ data sets of size $n = 200$ with a survival response Y_i, a single covariate X_i subject to a DL, and a vector of fully observed covariates $\boldsymbol{Z}_i$. Specifically, we generated the covariates $Z_{i1} \overset{iid}{\sim} \text{beta}(3,2) \cdot 84 + 18$, to represent an "age" variable, and $Z_{i2} \overset{iid}{\sim} \text{Bernoulli}(0.49)$, to represent a gender variable with $P(\text{male}) = 0.49$. We also generated a continuous covariate $Z_{i3}|Z_{i1}, Z_{i2} \overset{ind.}{\sim} \text{N}(\mu_i = 30 + 5Z_{i1} + 0.4Z_{i2}, \sigma^2 = 20)$. For the first set of simulations, we let $X_i|\boldsymbol{Z}_i \overset{ind.}{\sim} p\text{N}(\mu_i = (1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{\alpha}_1, \sigma^2 = 1) + (1-p)\text{N}(\mu_i = (1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{\alpha}_2, \sigma^2 = 1)$ where $p \overset{iid}{\sim} \text{Bernoulli}(0.3)$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T)^T = (5, -0.03, 1, 0.05, 2, 0.08, 1, -0.1)^T$. Values for the DL d, which we did not vary across individuals, were chosen so as to produce either 20% or 40% censoring on the covariate X_i.

The values for $\boldsymbol{\alpha}$ were chosen so that the distribution of $\boldsymbol{X_i}|\boldsymbol{Z}_i$ is clearly non-normal for many values of $\boldsymbol{Z}_i$ but the marginal distribution of X_i appears plausibly normal. Figure 1 shows a histogram, estimated density, and normal probability plot for 10,000 random samples from the distribution of X_i when there is no censoring as well as a histogram for 10,000 random samples from the distribution $X_i|\boldsymbol{Z_i}$ for a randomly generated $\boldsymbol{Z}_i$. While the marginal distribution of X_i appears reasonably normal, assuming normality for the conditional distribution of X_i would not be reasonable. We chose this set-up to demonstrate that it may be wise to choose a flexible distribution for $X_i|\boldsymbol{Z_i}$ even if the covariate appears to be marginally distributed according to a normal or another well-known density.

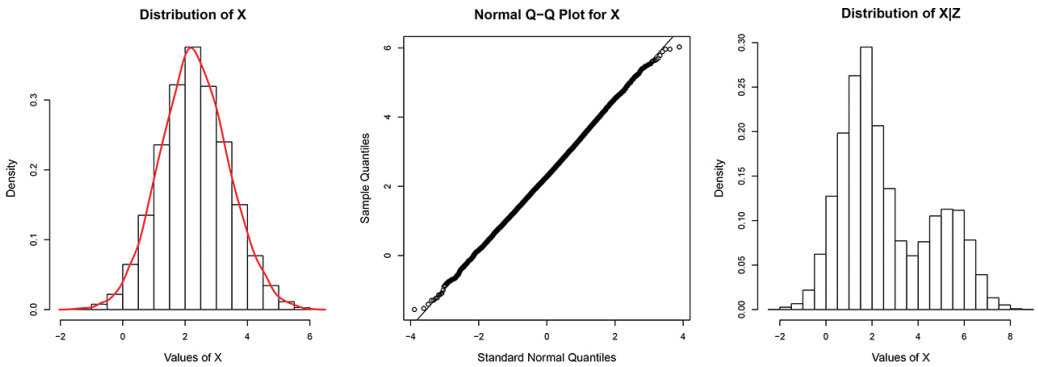

Figure 1: From left to right: (1) Histogram and estimated density for X based on 10,000 randoms samples; (2) Normal q-q plot based on same 10,000 random samples; (3) Histogram for X conditional on a single randomly generated $\boldsymbol{Z}$ based on 10,000 random samples.

We generated the log of the survival response as $\log(Y_i)|X_i, \boldsymbol{Z}_i \overset{ind.}{\sim} \text{TN}_{+\log(90)}(\mu_i = (1, X_i, \boldsymbol{Z}_i^T)\boldsymbol{\beta}, \sigma^2 = 4)$, where $\boldsymbol{\beta} = (2, 1, -0.01, -1, -0.05)^T$ and $\text{TN}_{+\log(90)}$ represents a normal distribution truncated above at log(90) since 90 days is the threshold in the GenIMS study. Finally, we let $Y_i = 90$ with probability $[1 + \exp\{(1, X_i, \boldsymbol{Z}_i^T)\boldsymbol{\gamma}\}]^{-1}$, where $\boldsymbol{\gamma} = (-1, 0.7, -0.05, -1, 0.03)^T$ so that about 35% of individuals were cured (i.e. $Y_i = 90$).

For the simulations in this section, we assumed that the distributional forms of the semicontinuous survival response were known in order to focus on assessing the proposed Monte Carlo EM estimation approach and the use of a mixture of normals to model the covariate subject to censoring. Specifically, we assumed it was known that a truncated normal AFT model describes the distribution of the log event times for those experiencing an event and a logistic regression describes the probability of never experiencing an event (surviving to 90 days). In Section 5.2, we study simulations for alternative time-to-event and binary regression scenarios.

We considered six data analysis approaches: our proposed semicontinuous survival method using a mixture of two normals for the distribution of $X_i|\boldsymbol{Z}_i$ (SS-MN), our proposed semicontinuous survival method using a normal to model the distribution of $X_i|\boldsymbol{Z}_i$ (SS-N), a multiple imputation method where 100 imputed data sets were generated assuming the semicontinuous survival model and a normal distribution to model $X_i|\boldsymbol{Z}_i$ (MI), an AFT and logistic regression model based on complete data where censored X_i values are replaced with DL/$\sqrt{2}$ (DL), an AFT and logistic regression model using only complete cases (CC), and an AFT and logistic regression model based on the omniscient knowledge of the uncensored covariate values (Omni).

For the MI approach, we used an improper imputation strategy based on fixed parameter estimates to generate imputations from $f(X_i|Y_i, \boldsymbol{Z}_i; \boldsymbol{\theta}) \propto f(Y_i|X_i, \boldsymbol{Z}_i; \boldsymbol{\beta}, \eta, \boldsymbol{\gamma}) f(X_i|Y_i, \boldsymbol{Z}_i; \boldsymbol{\theta}_x)$. Specifically, we used the complete case estimates for $\boldsymbol{\beta}$, η, and $\boldsymbol{\gamma}$ and maximum likelihood estimates for θ_x based on a censored normal model. Since these fixed estimates are consistent, the multiple imputation estimates will be asymptotically unbiased [46]. We included the multiple imputation estimator in the simulations for the purpose of comparison, though we do not believe it is an advantageous alternative to our proposed maximum likelihood approach for two reasons. First, obtaining imputations in a truncated region is not standard with current software and usual imputation strategies often do not work well with restricted ranges; thus, it is not computationally simpler. Second, it can be shown that this improper multiple imputation strategy is equivalent to a single iteration in an EM algorithm strategy for maximum likelihood [46, 47], which is the proposed approach.

Table 1: Single censored covariate: comparison of relative bias and standard deviations (in parenthesis below bias) of estimates for (a) the proposed semicontinuous survival model strategy assuming a mixture of normals distribution for the censored covariate (SS-MN) or a normal distribution for the censored covariate (SS-N); (b) the semicontinuous survival model using multiple imputations based on a normal distribution for the censored covariate (MI); (c) an AFT and logistic regression model replacing censored values with DL/$\sqrt{2}$ (DL); (d) an AFT and logistic regression model using only complete cases (CC); (e) an AFT and logistic regression model based on omniscient knowledge of the uncensored covariate data (Omni). Three underlying distributions for $X|Z$ are considered: a mixture of normals (Mixture), a normal distribution, and log-gamma. Estimates that are biased at the 0.05 level (without multiplicity corrections) are in bold.

Cens.	$X\|Z$	**Method**	β_0	β_1	β_2	β_3	β_4	γ_0	γ_1	γ_2	γ_3	γ_4
20%	Mixture	SS-MN	−0.01	−0.01	−0.03	0.01	0.00	0.03	**0.05**	0.03	**0.07**	**0.08**
			(1.13)	(0.19)	(0.01)	(0.43)	(0.01)	(1.13)	(0.17)	(0.02)	(0.50)	(0.01)
		SS-N	−0.02	0.00	−0.04	0.01	0.00	**0.04**	**0.07**	**0.03**	**0.07**	**0.08**
			(1.13)	(0.19)	(0.01)	(0.43)	(0.01)	(1.13)	(0.18)	(0.02)	(0.50)	(0.01)
		MI	−0.02	0.00	−0.04	0.01	0.00	**0.04**	**0.07**	**0.03**	**0.07**	**0.08**

			(1.14)	(0.19)	(0.01)	(0.43)	(0.01)	(1.13)	(0.18)	(0.02)	(0.50)	(0.01)
		DL	−0.04	0.01	−0.03	0.01	0.00	**0.05**	**0.08**	0.02	**0.06**	**0.07**
			(1.14)	(0.19)	(0.01)	(0.43)	(0.01)	(1.13)	(0.18)	(0.02)	(0.50)	(0.01)
		CC	0.00	−0.01	−0.02	0.00	−0.01	0.02	**0.05**	**0.03**	**0.09**	**0.07**
			(1.17)	(0.20)	(0.01)	(0.44)	(0.01)	(1.20)	(0.20)	(0.02)	(0.52)	(0.01)
		Omni	−0.01	−0.01	−0.03	0.01	−0.01	0.02	**0.04**	0.03	**0.08**	**0.07**
			(1.13)	(0.19)	(0.01)	(0.43)	(0.01)	(1.13)	(0.17)	(0.02)	(0.50)	(0.01)
40%	Mixture	SS-MN	0.02	−0.01	0.10	−0.03	−0.02	**0.06**	**0.08**	0.02	0.03	0.01
			(1.39)	(0.13)	(0.02)	(0.50)	(0.01)	(1.19)	(0.15)	(0.02)	(0.51)	(0.02)
		SS-N	**0.08**	−0.05	**−0.15**	**−0.05**	**0.03**	−0.03	−0.01	**0.25**	**0.14**	**0.47**
			(1.39)	(0.13)	(0.02)	(0.50)	(0.01)	(1.21)	(0.12)	(0.02)	(0.52)	(0.02)
		MI	0.06	**−0.03**	**−0.15**	**−0.05**	**0.04**	0.01	**0.03**	**0.24**	**0.16**	**0.40**
			(1.42)	(0.14)	(0.02)	(0.50)	(0.01)	(1.31)	(0.15)	(0.02)	(0.60)	(0.02)
		DL	−0.06	**0.08**	**−0.42**	**−0.04**	**0.19**	**0.19**	**0.24**	**0.38**	**0.21**	**0.79**
			(1.44)	(0.15)	(0.02)	(0.52)	(0.02)	(1.17)	(0.15)	(0.02)	(0.48)	(0.02)
		CC	0.04	−0.01	0.15	−0.02	−0.03	**0.14**	**0.17**	**0.09**	**0.28**	**0.10**
			(1.74)	(0.21)	(0.02)	(0.59)	(0.02)	(2.17)	(0.31)	(0.03)	(1.76)	(0.03)
		Omni	0.02	−0.01	0.10	−0.02	−0.02	**0.06**	**0.08**	0.03	0.02	0.02
			(1.34)	(0.13)	(0.02)	(0.49)	(0.01)	(1.16)	(0.14)	(0.02)	(0.49)	(0.02)
40%	Normal	SS-MN	−0.05	0.01	−0.01	−0.01	−0.02	**0.05**	**0.05**	**0.03**	**0.05**	0.03
			(1.17)	(0.16)	(0.01)	(0.46)	(0.01)	(1.02)	(0.15)	(0.01)	(0.41)	(0.01)
		SS-N	−0.05	0.00	−0.02	−0.02	−0.02	**0.05**	**0.05**	**0.04**	**0.05**	**0.05**
			(1.17)	(0.16)	(0.01)	(0.46)	(0.01)	(1.01)	(0.15)	(0.01)	(0.41)	(0.01)
		MI	**−0.07**	0.01	−0.03	−0.01	−0.01	**0.06**	**0.07**	**0.06**	**0.05**	**0.07**
			(1.20)	(0.16)	(0.01)	(0.46)	(0.01)	(1.06)	(0.16)	(0.02)	(0.42)	(0.01)
		DL	**−0.20**	**0.10**	**0.53**	−0.03	**−0.10**	**0.24**	**0.20**	**−0.11**	**0.10**	**−0.25**
			(1.22)	(0.18)	(0.01)	(0.47)	(0.01)	(0.98)	(0.17)	(0.01)	(0.38)	(0.01)
		CC	−0.06	0.01	−0.01	0.00	−0.02	**0.11**	**0.10**	**0.09**	**0.09**	**0.09**
			(1.56)	(0.23)	(0.02)	(0.54)	(0.02)	(1.62)	(0.25)	(0.02)	(0.60)	(0.02)
		Omni	−0.04	0.00	−0.00	−0.02	−0.02	**0.03**	**0.04**	**0.04**	**0.04**	**0.04**
			(1.14)	(0.15)	(0.01)	(0.46)	(0.01)	(0.93)	(0.12)	(0.01)	(0.40)	(0.01)
40%	LogGam	SS-MN	**−0.07**	0.03	−0.01	0.00	0.00	**0.07**	**0.07**	0.00	**0.05**	−0.03
			(1.18)	(0.24)	(0.02)	(0.46)	(0.02)	(1.02)	(0.20)	(0.02)	(0.37)	(0.01)
		SS-N	**−0.08**	**0.07**	**−0.24**	**0.06**	**0.07**	**0.10**	**0.17**	**0.08**	−0.02	**0.18**
			(1.19)	(0.24)	(0.02)	(0.46)	(0.02)	(1.00)	(0.21)	(0.01)	(0.37)	(0.01)
		MI	**−0.08**	**0.07**	**−0.20**	**0.06**	**0.07**	**0.11**	**0.17**	**0.07**	−0.01	**0.16**
			(1.22)	(0.25)	(0.02)	(0.46)	(0.02)	(1.03)	(0.23)	(0.02)	(0.38)	(0.01)
		DL	**−0.12**	0.02	**0.42**	**−0.09**	**−0.11**	**0.17**	**0.11**	**−0.09**	**0.13**	**−0.24**
			(1.21)	(0.24)	(0.02)	(0.46)	(0.01)	(0.97)	(0.20)	(0.01)	(0.35)	(0.01)
		CC	−0.02	0.00	−0.05	0.00	0.01	**0.08**	**0.08**	**0.08**	**0.13**	**0.08**
			(1.56)	(0.34)	(0.02)	(0.55)	(0.02)	(1.68)	(0.38)	(0.02)	(0.56)	(0.02)
		Omni	−0.04	0.00	−0.11	−0.03	0.01	**0.05**	**0.05**	**0.04**	**0.08**	**0.05**
			(1.10)	(0.20)	(0.01)	(0.43)	(0.01)	(0.94)	(0.15)	(0.01)	(0.36)	(0.01)

Table 1 displays the relative bias and standard deviation (in parentheses) of the parameter estimates for each of the six analysis methods. The relative bias for a parameter θ is estimated as

$$\frac{\hat{\theta} - \theta_0}{\theta_0},$$

where $\hat{\theta} = \sum_{j=1}^{N} \hat{\theta}_j / N$, $\hat{\theta}_j$ is the parameter estimate for the jth simulated data set, and θ_0 is the true parameter value. The standard deviations of the parameter estimator is estimated as

$$\sqrt{\frac{\sum_{j=1}^{N} (\hat{\theta}_j - \hat{\theta})^2}{N-1}}.$$

Relative biases that are statistically significant at the 0.05 level, without multiplicity corrections, are shown in bold. The first two sections in Table 1 give simulation results for two different censoring levels on X_i when the distribution of $X_i|\mathbf{Z}_i$ is actually a mixture of normals as described previously in the simulation set-up. The

last two sections display results when the actual distribution of the censored covariate is normal or log-gamma and centered at $(1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{\alpha}_1$, with an appropriate shift to preserve the 35% cure rate.

Due to the relatively small sample size for the simulations, there are minimal but statistically significant biases for several of the $\boldsymbol{\gamma}$ parameter estimates even when using the true data via the Omni analysis. The proposed SS-MN model performs comparably to the Omni analysis, with only slight losses in efficiency due to the covariate censoring, and significantly better than the other competing methods in terms of relative bias. Estimation using the SS-N method is reasonable only when the true distribution of the censored covariate is normal or the censoring percentage is low. While the relative biases for the MI method are similar to those for the SS-N method, the standard deviations of the γ_1 and γ_4 estimates are 10-20% higher while the standard deviations for the β_1 estimate are 3-6% higher. The DL strategy has substantially higher relative biases than the other methods, while the CC method leads to parameter estimates with standard deviations between 20% and 235% higher than the SS-MN method.

5.2 Single censored covariate: response model unknown or misspecified

In practice, the underlying distributions of the variables are usually unknown. In this section, we consider several scenarios with a single censored covariate where the parametric models in eq. (1) are unknown. In Section 5.2.1, we explore cases when the true distribution of event times for those experiencing the event is not lognormal. In Section 5.2.2, we consider a case where all of the parametric models are unknown - the AFT model, the binary response model, and the model for the covariates. Various flexible parametric models are assumed and are shown to perform well.

5.2.1 Misspecified AFT model

In the following simulations, we consider cases where the distribution of $X|\boldsymbol{Z}$ is a mixture of normals as defined in eq. (3), but the AFT model error term is incorrectly specified. Specifically, we generated the time-to-event data with the same location $(1, X_i, \boldsymbol{Z}_i^T)\boldsymbol{\beta}$ as defined in the Section 5.1, but with three different truncated event time distributions: a generalized gamma with scale 0.5 and shape 2, a generalized extreme value distribution with scale 3 and shape 0.05, and a mixture of a generalized gamma with scale 0.5 and shape 1 and a lognormal with scale 1.5. These three cases were considered because they represent a right-skewed log-survival response, a left-skewed log-survival response, and an unusually-shaped log-survival response.

We considered seven data analysis approaches, including the six previously described in Section 5.1. The additional approach we considered here assumes a truncated generalized gamma distribution for the time-to-event and a mixture of normal regressions for the distribution of $X_i|\boldsymbol{Z}_i$ (SS-MN-G). The generalized gamma distribution is a very flexible distribution that can take a variety of shapes, and it includes the exponential, Weibull, gamma, and lognormal distributions as special cases [48].

Table 2: Single censored covariate, misspecified AFT response model: comparison of relative bias and standard deviations (in parenthesis below bias) of estimates for (a) the proposed semicontinuous survival model strategy assuming a mixture of normals distribution for the censored covariate and a generalized gamma AFT (SS-MN-G); (b) the proposed semicontinuous survival model strategy assuming a mixture of normals distribution for the censored covariate (SS-MN-N) or a normal distribution for the censored covariate (SS-N) and a lognormal AFT; (c) the semicontinuous survival model using multiple imputations based on a normal distribution for the censored covariate and a lognormal AFT (MI); (c) a lognormal AFT and logistic regression model replacing censored values with DL/$\sqrt{2}$ (DL); (d) a lognormal AFT and logistic regression model using only complete cases (CC); (e) a lognormal AFT and logistic regression model based on omniscient knowledge of the uncensored covariate data (Omni). Estimates that are biased at the 0.05 level (without multiplicity corrections) are in bold.

$Y\|Z$	**Method**	β_0	β_1	β_2	β_3	β_4	γ_0	γ_1	γ_2	γ_3	γ_4
GenGam	SS-MN-G	0.03	−0.00	−0.02	−0.01	0.00	0.09	0.11	0.06	0.07	0.08
		(0.48)	(0.05)	(0.01)	(0.14)	(0.01)	(1.19)	(0.16)	(0.02)	(0.51)	(0.02)
	SS-MN-N	**−0.41**	0.03	−0.07	0.03	−0.02	0.09	0.11	0.05	0.06	0.03
		(0.78)	(0.09)	(0.01)	(0.29)	(0.01)	(1.19)	(0.17)	(0.02)	(0.52)	(0.02)
	SS-N	**−0.36**	0.00	−0.19	0.02	0.00	0.02	0.02	0.22	0.14	0.38
		(0.78)	(0.09)	(0.01)	(0.29)	(0.01)	(1.18)	(0.13)	(0.02)	(0.52)	(0.02)
	MI	**−0.37**	0.01	−0.19	0.02	0.00	0.04	0.04	0.20	0.14	0.29
		(0.79)	(0.10)	(0.01)	(0.29)	(0.01)	(1.25)	(0.16)	(0.02)	(0.55)	(0.02)
	DL	**−0.52**	0.10	−0.49	0.03	0.12	**0.17**	0.22	0.34	0.21	0.73

		(0.82)	(0.10)	(0.01)	(0.29)	(0.01)	(1.14)	(0.16)	(0.02)	(0.49)	(0.02)
	CC	**−0.33**	0.03	0.11	0.07	−0.03	0.15	0.17	0.08	0.14	0.07
		(1.00)	(0.14)	(0.01)	(0.34)	(0.01)	(1.86)	(0.28)	(0.03)	(0.84)	(0.03)
	Omni	**−0.40**	0.02	−0.03	0.04	−0.03	0.06	0.08	0.03	0.04	0.02
		(0.73)	(0.09)	(0.01)	(0.28)	(0.01)	(1.13)	(0.14)	(0.02)	(0.51)	(0.02)
GEV	SS-MN-G	−0.21	0.02	−0.23	−0.05	0.04	0.13	0.06	0.04	0.06	0.05
		(1.63)	(0.20)	(0.02)	(0.61)	(0.02)	(0.98)	(0.12)	(0.02)	(0.47)	(0.01)
	SS-MN-N	0.18	0.10	0.73	0.05	−0.24	0.14	0.07	0.04	0.07	0.04
		(1.63)	(0.26)	(0.03)	(0.69)	(0.03)	(0.98)	(0.12)	(0.02)	(0.47)	(0.01)
	SS-N	**0.67**	**−0.27**	−2.46	−0.28	0.55	0.07	0.03	−0.06	0.02	−0.16
		(1.60)	(0.19)	(0.03)	(0.68)	(0.03)	(0.98)	(0.12)	(0.02)	(0.47)	(0.01)
	MI	**0.36**	−0.08	−0.50	−0.09	0.04	0.07	0.03	−0.04	0.02	−0.13
		(1.59)	(0.21)	(0.02)	(0.63)	(0.03)	(1.02)	(0.13)	(0.02)	(0.47)	(0.01)
	DL	0.14	0.05	−4.16	**−0.51**	1.03	**0.74**	0.14	−0.23	−0.06	−0.44
		(1.75)	(0.30)	(0.02)	(0.67)	(0.02)	(0.97)	(0.12)	(0.01)	(0.45)	(0.01)
	CC	0.11	−0.07	0.26	−0.09	−0.13	0.12	0.07	0.07	0.09	0.10
		(2.10)	(0.34)	(0.03)	(0.80)	(0.03)	(1.34)	(0.18)	(0.02)	(0.55)	(0.02)
	Omni	**0.38**	−0.00	0.44	0.01	−0.15	0.04	0.05	0.06	0.07	0.09
		(1.53)	(0.23)	(0.03)	(0.68)	(0.03)	(0.97)	(0.12)	(0.02)	(0.47)	(0.01)
Mixture	SS-MN-G	−0.11	0.02	−0.22	−0.03	0.05	0.12	0.08	0.05	0.07	0.04
		(0.66)	(0.10)	(0.01)	(0.29)	(0.01)	(0.90)	(0.12)	(0.02)	(0.42)	(0.01)
	SS-MN-N	−0.13	0.02	−0.22	−0.04	0.04	0.13	0.09	0.06	0.07	0.05
		(0.67)	(0.10)	(0.01)	(0.30)	(0.01)	(1.01)	(0.13)	(0.02)	(0.45)	(0.01)
	SS-N	**0.18**	**−0.22**	−2.69	**−0.28**	0.66	0.16	0.08	−0.04	0.04	−0.14
		(0.73)	(0.12)	(0.01)	(0.31)	(0.01)	(1.01)	(0.13)	(0.02)	(0.44)	(0.01)
	MI	0.07	−0.14	−1.31	−0.15	0.30	0.09	0.07	0.02	0.06	−0.04
		(0.76)	(0.12)	(0.01)	(0.31)	(0.01)	(1.03)	(0.13)	(0.02)	(0.45)	(0.01)
	DL	**−0.32**	0.02	**−4.56**	**−0.52**	**1.15**	**0.58**	0.15	−0.19	−0.05	−0.48
		(0.77)	(0.14)	(0.01)	(0.31)	(0.01)	(0.97)	(0.12)	(0.01)	(0.43)	(0.01)
	CC	−0.12	−0.01	−0.06	−0.02	−0.03	0.10	0.09	0.09	0.08	0.09
		(1.08)	(0.17)	(0.02)	(0.40)	(0.02)	(1.24)	(0.17)	(0.02)	(0.53)	(0.02)
	Omni	−0.07	0.00	0.01	−0.01	−0.01	0.07	0.07	0.07	0.08	0.08
		(0.55)	(0.09)	(0.01)	(0.25)	(0.01)	(0.99)	(0.13)	(0.02)	(0.44)	(0.01)

Table 2 displays the relative bias and standard deviation (in parentheses) of the parameter estimates for each of the analysis methods, with relative bias estimates in bold if the absolute bias is different from 0 at the 0.05 level. Overall, the SS-MN-G and SS-MN-N methods perform better than the SS-N and DL methods in terms of bias and better than the CC and MI methods in terms of efficiency. The SS-MN-G method additionally shows a potential improvement in bias reduction compared to the Omni and CC methods since they incorrectly assume a lognormal event time distribution. In all cases, the proposed SS-MN-G method performs well. For this reason, we suggest generally approaching modeling eq. (2) using a flexible AFT distribution such as the generalized gamma.

5.2.2 All parametric models unknown or misspecified

We now consider a scenario where all of the parametric models are unknown. Specifically, we let the conditional covariate distribution be log-gamma as defined in Section 5.1, the AFT error model be a mixture of generalized gamma and lognormal distributions as defined in Section 5.2.2, and the binary regression model be probit rather than logistic. The same seven methods are considered as in Section 5.2.2, but AIC was used to determine whether the binary submodel should be logistic or probit.

Table 3 displays the relative bias and standard deviation (in parentheses) of the parameter estimates for each of the analysis methods, with relative bias estimates in bold if the absolute bias is different from 0 at the 0.05 level. As with the simulations in the previous section, the SS-MN-G and SS-MN-N methods perform well, will lower bias than the SS-N and DL methods and better efficiency than the CC method. Interestingly, the MI method is fairly comparable in terms of bias and efficiency to the SS-MN-G and SS-MN-N methods in this particular scenario, though it was shown in other cases to be inferior.

Table 3: Single censored covariate, all parametric models unknown: comparison of relative bias and standard deviations (in parenthesis below bias) of estimates for (a) the proposed semicontinuous survival model strategy assuming a mixture of normals distribution for the censored covariate and a generalized gamma AFT (SS-MN-G); (b) the proposed semicontinuous survival model strategy assuming a mixture of normals distribution for the censored covariate (SS-MN-N) or a normal distribution for the censored covariate (SS-N) and a lognormal AFT; (c) the semicontinuous survival model using multiple imputations based on a normal distribution for the censored covariate and a lognormal AFT (MI); (c) a lognormal AFT and logistic regression model replacing censored values with DL/$\sqrt{2}$ (DL); (d) a lognormal AFT and logistic regression model using only complete cases (CC); (e) a lognormal AFT and logistic regression model based on omniscient knowledge of the uncensored covariate data (Omni). Estimates that are biased at the 0.05 level (without multiplicity corrections) are in bold.

Method	β_0	β_1	β_2	β_3	β_4	γ_0	γ_1	γ_2	γ_3	γ_4
SS-MN-G	−0.20	0.11	0.06	0.03	0.01	**0.75**	0.07	0.03	0.06	0.10
	(0.92)	(0.18)	(0.01)	(0.34)	(0.01)	(0.89)	(0.20)	(0.02)	(0.34)	(0.02)
SS-MN-N	−0.17	0.08	0.27	0.04	−0.02	**0.75**	0.05	0.07	0.05	0.08
	(0.83)	(0.16)	(0.01)	(0.33)	(0.01)	(0.79)	(0.18)	(0.01)	(0.33)	(0.01)
SS-N	**−0.47**	**0.24**	−0.33	0.15	0.16	**0.96**	0.12	0.03	0.07	−0.00
	(0.85)	(0.18)	(0.01)	(0.33)	(0.01)	(0.78)	(0.18)	(0.01)	(0.33)	(0.01)
MI	**−0.32**	**0.17**	−0.31	0.15	0.15	**0.82**	0.07	0.04	0.05	0.02
	(0.89)	(0.17)	(0.01)	(0.33)	(0.01)	(0.78)	(0.17)	(0.01)	(0.33)	(0.01)
DL	**−0.30**	0.09	0.66	−0.18	−0.16	**0.48**	−0.04	0.05	0.02	0.08
	(0.86)	(0.17)	(0.01)	(0.35)	(0.01)	(0.69)	(0.15)	(0.01)	(0.32)	(0.01)
CC	−0.11	0.01	0.08	0.02	−0.03	**0.84**	0.10	0.08	0.08	0.08
	(1.43)	(0.30)	(0.01)	(0.44)	(0.01)	(1.25)	(0.25)	(0.01)	(0.36)	(0.01)
Omni	−0.05	−0.01	0.05	−0.01	-0.01	**0.81**	0.08	0.05	0.07	0.04
	(0.63)	(0.09)	(0.01)	(0.26)	(0.01)	(0.76)	(0.17)	(0.01)	(0.33)	(0.01)

5.3 Multiple censored covariates

We conducted simulations similar to those in Section 5.1, but with three covariates subject to censoring below DLs so as to mimic the GenIMS application more closely. Specifically, for $N = 300$ data sets of size $n = 200$, we generated three fully observed covariates Z_{i1}, Z_{i2}, and Z_{i3} in the same way as in Section 5.1. We then considered two strategies for simulating $(X_{i1}, X_{i2}, X_{i3})|\boldsymbol{Z}_i$. First, we let $(X_{i1}, X_{i2}, X_{i3})|\boldsymbol{Z}_i \overset{ind.}{\sim} p\mathrm{N}_3(\boldsymbol{\mu}_i^T = (1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{A}_1, \Sigma) + (1-p)\mathrm{N}_3(\boldsymbol{\mu}_i^T = (1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{A}_2, \Sigma)$ where $p \overset{iid}{\sim} \mathrm{Bernoulli}(0.3)$, N_3 is a three-variate normal distribution, and

$$\boldsymbol{A}_1 = \begin{pmatrix} 3.0 & 0.05 & -0.03 & 1.0 \\ 2.5 & -0.10 & 0.05 & 1.2 \\ 1.2 & 0.04 & 0.09 & -1.1 \end{pmatrix}^T, \quad \boldsymbol{A}_2 = \begin{pmatrix} 1.5 & 0.03 & -0.07 & 1.5 \\ 2.0 & -0.05 & 0.05 & 0.4 \\ 1.5 & 0.04 & 0.03 & -0.5 \end{pmatrix}^T, \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.5 & 0.5 \\ 0.5 & 1.0 & 0.5 \\ 0.5 & 0.5 & 1.0 \end{pmatrix}.$$

These parameter values were chosen so that the conditional distribution of $\boldsymbol{X}_i$ was somewhat skewed and all of the covariates were reasonably correlated. Second, we generated $(X_{i1}, X_{i2}, X_{i3})|\boldsymbol{Z}_i$ using a Gaussian copula with mean $(1, Z_{i1}, Z_{i2}, Z_{i3})\boldsymbol{A}_1$, covariance Σ, and marginal distributions gamma(3,1), N(0,1), and t(df = 5). This distribution was chosen in order to check the performance of the multivariate mixture of normals approach for a somewhat unusual joint distribution for the censored covariates. Values for the vector of DLs $\boldsymbol{d}$, which we did not vary across individuals, were chosen so that 20%, 50%, and 40% of the observations for X_{i1}, X_{i2}, and X_{i3}, respectively, were censored.

The log-survival response was generated as in Section 5.1, except with the parameter vector $\boldsymbol{\beta} = (8, -0.5, -0.2, -0.2, -0.03, -0.01, -1)^T$. Also, we let $Y_i = 90$ with probability $[1 + \exp\{(1, \boldsymbol{X}_i, \boldsymbol{Z}_i^T)\boldsymbol{\gamma}\}]^{-1}$ and $\boldsymbol{\gamma} = (1.5, -0.05, -0.06, -0.03, -0.02, -0.005, -0.5)^T$, so that about 35% of individuals were cured (i.e. $Y_i = 90$). Again, we used a normal AFT model to describe the distribution of event times for those experiencing an event and a logistic regression to model the probability of never experiencing an event (surviving to 90 days).

Table 4: Multiple censored covariates: comparison of relative bias and standard deviations of estimates for (a) the proposed semicontinuous survival model strategy assuming a series of mixture of normal distributions for the censored covariates (SS-MN) or a series of normal distributions for the censored covariates (SS-N); (b) the semicontinuous survival model using multiple imputations based on a series of normal distributions for the censored covariates (MI); (c) an AFT and logistic regression model replacing censored values with DL/$\sqrt{2}$ (DL); (d) an AFT and logistic regression model using only complete cases (CC); (e) an AFT and logistic regression model based on the omniscient knowledge of the true uncensored covariate data (Omni). Two underlying distributions for $\boldsymbol{X}|Z$ are considered: a mixture of normals (Mixture) and a Gaussian copula with gamma, normal, and t marginal distributions. Estimates that are biased at the 0.05 level (without multiplicity corrections) are in bold.

		Method											
		SS-MN		SS-N		MI		DL		CC		Omni	
$X\|Z$	Par.	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
Mixture	β_0	0.00	1.70	**−0.04**	1.69	0.00	1.75	**−0.15**	1.46	−0.07	3.91	−0.02	1.52
	β_1	0.00	0.22	−0.03	0.22	**−0.07**	0.23	**−0.23**	0.19	−0.05	0.53	−0.02	0.21
	β_2	**0.24**	0.30	0.08	0.29	0.15	0.31	**−1.66**	0.31	−0.24	0.67	−0.06	0.18
	β_3	−0.24	0.33	**−0.35**	0.35	−0.01	0.35	**0.82**	0.23	−0.15	0.61	−0.04	0.27
	β_4	0.01	0.03	−0.03	0.03	**−0.13**	0.03	**−0.74**	0.01	−0.09	0.03	−0.04	0.02
	β_5	0.09	0.03	0.46	0.04	−0.14	0.04	−0.31	0.02	0.18	0.06	0.15	0.03
	β_6	−0.01	0.73	−0.02	0.74	**0.15**	0.80	**0.54**	0.66	−0.03	1.64	−0.02	0.66
	γ_0	0.03	0.86	0.02	0.82	0.09	0.96	**−0.49**	0.79	**0.46**	3.44	0.05	0.80
	γ_1	−0.35	0.17	−0.37	0.16	−0.38	0.20	**−1.33**	0.14	0.06	0.49	−0.26	0.14
	γ_2	−0.28	0.19	−0.36	0.18	0.10	0.25	**−4.07**	0.19	−0.13	0.63	−0.07	0.13
	γ_3	0.25	0.20	0.09	0.22	0.47	0.24	**2.88**	0.16	1.78	0.57	0.51	0.18
	γ_4	−0.02	0.01	−0.03	0.01	0.01	0.01	**−0.86**	0.01	0.28	0.04	0.02	0.01
	γ_5	0.13	0.02	0.26	0.02	−0.05	0.02	**1.23**	0.02	0.16	0.06	−0.06	0.02
	γ_6	**0.17**	0.53	0.15	0.55	**0.20**	0.55	**0.52**	0.42	**0.43**	1.41	0.05	0.46
Copula	β_0	**−0.07**	1.78	**−0.08**	1.80	**−0.06**	2.07	**−0.19**	2.53	−0.05	3.66	**−0.05**	1.72
	β_1	**−0.06**	0.15	**−0.07**	0.15	−0.04	0.15	0.03	0.17	−0.05	0.23	−0.02	0.15
	β_2	−0.18	0.29	−0.19	0.29	−0.16	0.34	**−0.74**	0.49	−0.01	0.59	−0.10	0.27
	β_3	**−0.18**	0.23	**−0.19**	0.23	−0.03	0.25	−0.20	0.30	−0.01	0.37	−0.15	0.21
	β_4	−0.05	0.01	−0.05	0.01	**−0.07**	0.01	−0.74	0.01	−0.07	0.02	**−0.07**	0.01
	β_5	−0.15	0.02	−0.14	0.02	−0.19	0.02	**0.81**	0.02	−0.13	0.02	−0.16	0.02
	β_6	0.02	0.49	−0.02	0.49	−0.02	0.51	**0.11**	0.49	−0.07	0.72	−0.01	0.47
	γ_0	0.07	1.24	0.07	1.23	0.10	1.54	**−0.56**	1.89	0.31	3.26	0.06	1.14
	γ_1	0.17	0.12	0.13	0.12	0.22	0.12	0.11	0.13	0.47	0.21	0.33	0.12
	γ_2	0.01	0.24	0.03	0.24	0.13	0.29	**−1.50**	0.40	0.60	0.57	−0.27	0.22
	γ_3	−0.75	0.18	−0.77	0.18	−0.42	0.22	**−2.30**	0.25	−0.32	0.43	−0.37	0.16
	γ_4	**0.07**	0.01	**0.07**	0.01	**0.07**	0.01	**−0.69**	0.01	**0.23**	0.02	0.07	0.01
	γ_5	0.11	0.01	0.11	0.01	0.12	0.01	**1.61**	0.01	0.23	0.02	0.16	0.01
	γ_6	0.01	0.35	0.02	0.35	0.01	0.37	**0.20**	0.33	−0.04	0.69	0.04	0.33

Table 4 displays the relative bias and standard deviation of parameter estimates for the same six data analysis approaches as in Section 5.1: our proposed semicontinuous survival method using a mixture of two normals for modeling each of the univariate distributions in eq. (4) (SS-MN), our proposed semicontinuous survival method using a normal distribution to model each of the univariate distributions in eq. (4) (SS-N), a multiple imputation method where 100 imputed data sets were generated based on the semicontinuous survival model and a normal distribution to model each of the univariate distributions in eq. (4), an AFT and logistic regression model based on complete data where censored $\boldsymbol{X}_i$ values are replaced with DL/$\sqrt{2}$ (DL), an AFT and logistic regression model using only complete cases (CC), and an AFT and logistic regression model based on the omniscient knowledge of the uncensored covariate values (Omni). Additionally, due to the high censoring percentages for X_{i2} and X_{i3}, we considered a hybrid method, which, based on eq. (4), assumed a mixture of normal distributions for modeling the conditional distribution of $X_{i1}|X_{i2}, X_{i3}, \boldsymbol{Z}_i$ and normal distributions for modeling the conditional distributions of $X_{i2}|X_{i3}, \boldsymbol{Z}_i$ and $X_{i3}|\boldsymbol{Z}_i$. However, the results for this method are not included because they were very similar to those for the SS-MN method.

The SS-MN, SS-N, and MI methods all perform similarly in terms of bias, with the CC method perhaps performing slightly better. However, the SS-MN and SS-N methods yield substantially more efficient estimates than the CC and MI methods, with standard deviations of the parameter estimates for the CC and MI methods as much as 280% and 31% higher, respectively. The mean squared errors (not shown) for the CC and MI method are significantly higher than the SS-MN and SS-N methods for all parameter estimates. The DL method estimates are poor on all accounts, with statistically significant biases for 23 of the 28 parameter estimates and magnitudes of relative bias as high as 288%. The results are similar for both the case when the distribution of the censored covariates is truly a mixture of normals and when it is generated using the copula, though there are slightly higher biases in the latter case.

We conducted an additional simulation for another unusual multivariate distribution of the censored covariates, with results included in the online Supplementary Material. The relative performance of all of methods remained the same.

6 Application

We illustrate the semicontinuous survival analysis method developed in Sections 2 and 3 by applying it to the Genetic and Inflammatory Markers of Sepsis (GenIMS) data set. One of the main purposes of the GenIMS study was to determine the relationship between cytokine levels and 90-day survival (event of surviving at least 90 days) for patients with community acquired pneumonia (CAP) [10]. It is additionally relevant to model the relationship between these cytokines and survival for those not surviving to 90 days. Thus, we would like to model survival times where a point mass occurs at 90 days.

Cytokines are cell-signaling protein molecules that are sent out by cells in the immune system. Three cytokines were measured in this study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). The TNF and IL-6 cytokines serve as biomarkers of pro-inflammatory responses to CAP while IL-10 serves as a biomarker of anti-inflammatory responses to CAP. It has been thought that pro- and anti-inflammatory responses in the body help explain the development of severe sepsis and resulting deaths, and that understanding these relationships could be important for developing medical treatments.

The data for the GenIMS study were obtained by first enrolling individuals with CAP immediately after admission to a hospital and then collecting biological measurements and demographic information for each individual. In our analysis, we considered the 1418 patients who actually acquired CAP, necessitated a hospital stay, and had TNF, IL-6, and IL-10 measurements taken on the first day of hospitalization.

We fit a semicontinuous survival model for the event times based on six covariates: sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), age, and TNF, IL-6, and IL-10 measurements on the first day of hospitalization. While sex, race, and age were fully observed, the cytokine biomarker measurements were censored below the detection thresholds 4, 2 or 5, and 5 pg/ml, respectively, with censoring proportions of 35.54%, 13.40%, and 46.83%, respectively. We also considered using Apache III scores (a measure of severity of the CAP) as a covariate, but found they were very collinear with the cytokine levels and decided to leave them out of the analysis.

We used a generalized gamma AFT model to explain the relationship between the event time and the covariates of interest and a logistic regression model to describe the relationship between 90-day survival and the covariates of interest. While a few previous works assumed a lognormal distribution for the covariates [10, 13], we assumed that conditional on the sex, race, and age covariates, TNF, IL-6, and IL-10 are distributed according to series of mixture of normal distributions as in eq. (4). We obtained parameter estimates for each of the models using the EM algorithm described in Section 3.

Table 5 displays the analysis results for the AFT and logistic submodels. For the AFT submodel, we observe that for those individuals who do not survive 90 days after admittance to the hospital, higher levels of the cytokine IL-6 are associated with shorter survival times, conditional on TNF and IL-10 levels, sex, race, and age (p-value = 0.002). All of the other variables of interest are not statistically significant, possibly due to the relatively low percentage of individuals dying before 90 days (11.7%). For the logistic submodel, we observe that the probability of surviving to day 90 is only significantly associated with sex and age, conditional on race and the three cytokines, with younger females being the most like to survive to the 90-day mark.

Table 5: Coefficient parameter estimates and standard errors for the semicontinuous survival model for survival time based on the covariates sex, race, age, and the three cytokine covariates of interest,TNF, IL-6, and IL-10.

AFT Model	**Intercept**	**Sex**	**Race**	**Age**	**TNF**	**IL-6**	**IL-10**
Estimate	11.985	−0.192	−0.092	−0.001	0.16	−0.285	−0.011
SE	1.42	0.40	0.75	0.01	0.21	0.09	0.12
p-value	< 0.001	0.630	0.903	0.922	0.434	0.002	0.930
Logistic Model	**Intercept**	**Sex**	**Race**	**Age**	**TNF**	**IL-6**	**IL-10**
Estimate	7.406	−0.402	−0.195	−0.059	−0.075	−0.080	−0.107
SE	0.78	0.18	0.31	0.01	0.11	0.06	0.07
p-value	< 0.001	0.024	0.530	< 0.001	0.482	0.179	0.103

For each of the three cytokines of interest, we conducted two degree of freedom χ^2 tests for $H_0 : \beta_{cytokine} = \gamma_{cytokine} = 0$. This type of test, based on both the AFT and logistic parts of the model, is often of main interest in semicontinuous models, and for the GenIMS study represents a test that each individual cytokine affects survival in any way, whether it be surviving longer/shorter or having a higher/lower chance of being cured of CAP. Only IL-6 was found to be important for predicting survival (p-value = 0.004); there is little evidence for

the importance of TNF (p-value = 0.585) and only moderate evidence for IL-10 (p-value = 0.117). A six degree of freedom χ^2 test for all of the cytokine parameters is only moderately significant (p-value = 0.093).

Due to a relatively high level of pairwise correlations among the three cytokines for the observations above the DLs (0.25-0.35), we also considered modeling each cytokine separately. In each of these models, we included sex, race, age, and one of the cytokines in both the AFT and logistic parts of the model.

Table 6 displays the results of the three single cytokine models. Unsurprisingly, the estimated coefficients for sex, race, and age are similar across all three single cytokine models and the multivariate model in Table 3. Surprisingly, perhaps, the estimated coefficients for the three different cytokines are also similar across the three separate models — in contrast to the multivariate model with results in Table 3 — indicating similar marginal associations of the cytokines with survival time. Standard errors for the estimates for the cytokine coefficients varied significantly across the three models, primarily due to different levels of censoring. There is strong evidence that, conditional on sex, race, and age, higher IL-6 levels are associated with a lower probability of surviving to 90 days and a shorter survival time. There is also strong evidence that IL-10 is conditionally related to surviving 90 days, but only moderate evidence that it is associated with survival time for those not surviving 90 days. The TNF cytokine is only moderately statistically significant in the logistic model. In all three models, there is a strong relationship between sex and age and survival to 90 days, with younger females being the most like to survive to the 90-day mark.

Table 6: Coefficient parameter estimates and standard errors for the semicontinuous survival model for three different sets of covariates: one of three cytokine covariates of interest together with the covariates gender, race, and age.

	TNF			**IL-6**			**IL-10**		
	Est.	**SE**	**p-value**	**Bias**	**SE**	**p-value**	**Bias**	**SE**	**p-value**
AFT Model									
Intercept	13.804	3.771	< 0.001	13.029	3.297	< 0.001	13.910	3.766	< 0.001
Cytokine	−0.141	0.169	0.407	−0.252	0.079	0.001	−0.167	0.112	0.136
Sex	−0.067	0.376	0.859	−0.203	0.323	0.531	−0.011	0.363	0.976
Race	−0.281	0.645	0.663	−0.132	0.557	0.812	0.007	0.640	0.991
Age	0.010	0.017	0.574	0.001	0.016	0.982	−0.010	0.016	0.541
Logistic Model									
Intercept	7.065	0.635	< 0.001	7.271	0.635	< 0.001	7.125	0.629	< 0.001
Cytokine	−0.188	0.091	0.039	−0.137	0.042	0.001	−0.178	0.054	< 0.001
Sex	−0.452	0.175	0.010	−0.422	0.175	0.016	−0.411	0.176	0.019
Race	−0.264	0.317	0.405	−0.210	0.315	0.506	−0.185	0.316	0.558
Age	−0.057	0.007	< 0.001	−0.058	0.007	< 0.001	−0.059	0.007	< 0.001

We conducted two degree of freedom χ^2 tests for $H_0 : \beta_{cytokine} = \gamma_{cytokine} = 0$ for all three models. In the IL-6 model and IL-10 models, the p-value for this test is < 0.001 and 0.001, respectively, while in the TNF model the p-value is 0.092. Thus, there is strong evidence that IL-6 and IL-10, conditional on sex, race, and age, are associated with decreased survival time and 90-day survival chances, but only moderate evidence of this relationship for TNF.

7 Discussion

We have proposed a semicontinuous survival model to handle time-to-event data with a point mass at a known cure threshold and one or more covariates subject to DLs. We additionally proposed a Monte Carlo EM algorithm for obtaining estimates in this model. Through simulations, we have shown that our proposed method flexibly models the distribution of the censored covariates and leads to approximately unbiased estimates for the parameters of interest in a variety of scenarios.

We note that while the methodology developed in this paper was motivated by a special case of long-term survival data where a cure time is known and no event times were censored prior to this time, the proposed model and estimation method could be applied to any semicontinuous regression model with a covariate censored due to DLs. It may also be noted that censoring due to DLs is just a special case of coarsened data and that a similar methodology could be applied to handle more general types of missing data in semicontinuous models, as has been done by authors in other regression contexts with missing covariates.

The methods proposed in this paper do have a few shortcomings. Perhaps foremost, it may be noted that the proposed methods seem specialized to a particular data set with a known cure threshold and no censoring before the threshold. However, in practice it has been relatively common to analyze survival data as a binary outcome [49]. For example, with various cancers and diseases, 2- and 5-year survival rates are important medical summary statistics. The proposed method has the advantage that it allows for simultaneously modeling survival probabilities at these times and the survival time for those not making it to the threshold of interest. If censoring occurs before the threshold of interest for some individuals, then the true survival time and cure status would be unknown. Bernhardt [50] considers this special case of cure models with a cure threshold in the context of fully observed covariates. To handle covariates subject to DLs, a similar Monte Carlo EM algorithm as that proposed in this paper could be developed in a cure model context. As another case where a semicontinuous survival model could apply, consider a timed sporting competition that includes a binary outcome. For example, in marathon qualifying events, we may be interested in a single model for the distribution of finishing times for qualifiers as well as the probability of qualifying.

A second disadvantage of our method is that it is somewhat computationally intensive. We proposed a Monte Carlo EM algorithm where each step requires obtaining many imputations and conducting three or more independent optimization routines. With relatively few parameters in the model and a single covariate subject to DLs, it may be faster to directly maximize the observed data likelihood using numerical integration techniques to approximate the integral, though numerical issues would be more likely to arise. However, we note that for the simulation and application data sets, computational effort was a relatively minor issue as the parameter estimates were obtained in a few minutes or less in the case of one covariate subject to a detection limit and less than 30 minutes with three covariates. If time was an issue, a simpler model than a mixture of normals for $\mathbf{X}_i|\mathbf{Z}_i$ would decrease computational time. Additionally, we emphasize that one main advantage of the proposed method is that it is straightforward and easily implemented in statistical programs. Alternatively, as shown in the simulations in Section 5, a multiple imputation approach that does not require iterative maximizations only sacrifices a small amount of efficiency in parameter estimation.

Acknowledgements:

The author would like to thank the editors and reviewers for their valuable comments. The author would also like to thank Dr. Lan Kong and the CRISMA (Clinical Research, Investigation, and Systems Modeling of Acute Illness) Center at the University of Pittsburgh for providing the GenIMS data set.

References

[1] Smith VE, Preisser JS, Neelon B, Maciejewski ML. A marginalized two-part model for semicontinuous data. Stat Med. 2014;33:4891–4903.
[2] Zhou X-H, Tu W. Comparison of several independent population means when their samples contain log-normal and possibly zero observations. Biometrics 1999;55:645–651.
[3] Liu L, Strawderman RL, Johnson BA, O'Quigley JM. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. Stat Meth Med Res. 2016;25:133–152.
[4] Mills ED. Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data. Ph.D. thesis, University of Iowa, 2013.
[5] Su L, Tom BD, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics 2009;10:374–389.
[6] Manning Jr. WG, Morris CN, Newhouse JP, Duan N, EB Keeler, Liebowitz A, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. Health, Econ Health Econ. 1983;1:103–123.
[7] Duan N, Manning Jr. WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. J Bus Econ Stat. 1983;1:115–126.
[8] Moulton LH, Curriero FC, Barroso PF. Mixture models for quantitative HIV RNA data. Stat Meth Med Res. 2002;11:317–325.
[9] Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. J Am Stat Assoc. 2001;96:730–745.
[10] Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, et al. Understanding the inflammatory cytokine response in pneumonia and sepsis. Archives of Internal Med. 2007;167:1655–1663.
[11] Angus DC, Carlet J. Surviving intensive care: a report from the 2002 brussels roundtable. Intensive Care Med. 2003;29:368–377.
[12] Cohen J, Guyatt G, Bernard GR, Calandra T, Cook D, Elbourne D, et al. New strategies for clinical trials in patients with sepsis and septic shock. Intesive Care Med. 2001;29:880–886.
[13] Bernhardt PW, Wang HJ, Zhang S. Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. Comput Stat Data Anal 2014;69:81–91.
[14] D' Angelo GD, Weissfeld L. An index approach for the cox model with left censored covariates. Stat Med. 2008;27:4502–4514.
[15] Sattar A, Sinha SK, Morris NJ. A parametric survival model when a covariate is subject to left-censoring. J Biomet Biostat. 2012;S3:002. DOI: 10.4172/2155–6180.S3–002.
[16] Berkson J, Gage RP. Survival cure for cancer patients following treatment. J Am Stat Assoc. 1952;47:501–515.

[17] Boag JW. Maxmimum likelihood estimates of the proportion of patients cured by cancer therapy. J R Stat Soc, Ser B. 1949;11:15–53.
[18] Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg. 1990;5:46–51.
[19] Austin PC, Hoch JS. Estimating linear regression models in the presence of a censored independent variable. Stat Med. 2004;23:411–429.
[20] Giovanini J. Generalized linear mixed models with censored covariates. Ph.D. thesis. Oregon State University, 2008.
[21] Austin PC, Brunner LJ. Type I error inflation in the presence of a ceiling effect. Am Statistician. 2003;57:97–104.
[22] Bernhardt PW, Wang HJ, Zhang S. Statistical methods for generalized linear models with covariates subject to detection limits. Stat Biosci. 2015;7:68–89.
[23] Helsel DR. Statistics for censored environmental data using minitab and R, 2nd ed. Wiley, 2012.
[24] Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004;112:1691–1696.
[25] Lynn HS. Maximum likelihood inference for left-censored hiv rna data. Stat Med. 2001;20:33–45.
[26] Rigobon R, Stoker TM. Estimation with censored regressors: basic issues. Int Econ Rev. 2007;48:1441–1467.
[27] Rigobon R, Stoker TM. Bias from censored regressors. J Bus Econ Stat. 2009;27:340–353.
[28] Langohr K, Gomez G, Muga R. A parametric survival model with an interval-censored covariate. Stat Med. 2004;23:309–319.
[29] Lee S, Park SH, Park J. The proportional hazards regression with a censored covariate. Stat Probab Lett. 2003;61:309–319.
[30] Chen Q, Wu H, Ware LB, Koyama T. A Bayesian approach for the Cox proportional hazards model with covariates subject to detection limit. Int J Stat Med Res. 2014;3:32–43.
[31] Sattar A, Sinha SK, Wang X-H, Li Y. Frailty models for pneumonia to death with a left-censored covariate. Stat Med. 2015;34:2266–2280.
[32] Quandt R, Ramsey J. Estimating mixutres of normal distributions and switching regression. J Am Stat Assoc. 1978;73:730–738.
[33] Norets A. Approximation of conditional densities by smooth mixtures of regression. Ann Stat. 2010;38:1733–1766.
[34] Ibrahim J. Monte Carlo EM for missing covariates in parametric regression models. Biometrics 1999;55:591–596.
[35] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc, Ser B. 1977;39:1–38.
[36] McLachlan GJ, Krishnan T. The EM Algorithm and Extensions, 2nd ed. John Wiley and Sons, Inc., 2008.
[37] Wei GC, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. J Am Stat Assoc. 1990;85:699–704.
[38] May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. Stat Med. 2011;30:2551–2561.
[39] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 1953;21:1087–1092.
[40] Faria S, Soromenho G. Fitting mixtures of linear regressions. J Stat Comput Simul. 2010;80:201–225.
[41] Jackson C. 2013 flexmix: flexible parameter survival and multi-slate models. Available at: http://CRAN.R-project.org/package=flexsurv, R package version 0.7.
[42] Geyer CJ. 2016 mcmc: Markov chain Monte Carlo. Available at: http://CRAN.R-project.org/package=mcmc, R package version 0.9-4.
[43] Louis TA. Finding the observed information matrix when using the EM algorithm. J R Stat Soc, Ser B. 1982;44:226–233.
[44] Gelman A, Mechelan IV, Verbeke G, Heitjan D, Meulders M. Multiple imputation for model checking: completed-data plots with missing and latent data. Biometrics 1977;61:74–85.
[45] Bernhardt PW. Model validation and influence diagnostics for regression models with missing covariates. Statistics in Medicine 2018. DOI: 10.1002/sim.7584.
[46] Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. Biometrika 1998;84:935–948.
[47] Tsiatis AA. Semiparametric Theory and Missing Data. Springer, 2006.
[48] Cox C, Chu H, Schneider MF, Mu noz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Stat Med. 2007;23:4352–4374.
[49] Lim J, Lee KE, Hahn KS, Park K. Analyzing survival data as binary outcomes with logistic regression. Commun Korean Stat Soc. 2010;17:117–126.
[50] Bernhardt PW. A flexible cure rate model with dependent censoring and a known cure threshold. Stat Med. 2016;25. DOI: 10.1002/sim.7014.

Supplementary Material: The online version of this article offers supplementary material (DOI:https://doi.org/10.1515/ijb-2017-0058).