Published in final edited form as:

Comput Stat Data Anal. 2014 January ; 69: . doi:10.1016/j.csda.2013.07.027.

Flexible Modeling of Survival Data with Covariates Subject to Detection Limits via Multiple Imputation

Paul W. Bernhardt^{1,*}, Huixia Judy Wang, and Daowen Zhang

Department of Statistics, North Carolina State University, Raleigh, NC, USA

Abstract

Models for survival data generally assume that covariates are fully observed. However, in medical studies it is not uncommon for biomarkers to be censored at known detection limits. A computationally-efficient multiple imputation procedure for modeling survival data with covariates subject to detection limits is proposed. This procedure is developed in the context of an accelerated failure time model with a flexible seminonparametric error distribution. The consistency and asymptotic normality of the multiple imputation estimator are established and a consistent variance estimator is provided. An iterative version of the proposed multiple imputation algorithm that approximates the EM algorithm for maximum likelihood is also suggested. Simulation studies demonstrate that the proposed multiple imputation methods work well while alternative methods lead to estimates that are either biased or more variable. The proposed methods are applied to analyze the dataset from a recently-conducted GenIMS study.

Keywords

Accelerated failure time model; Censored predictor; Complete case; Detection limit; Multiple imputation; Seminonparametric distribution

1. Introduction

Biomedical datasets frequently contain variables that are subject to censoring. Censoring due to detection limits (DLs) often occurs in practice when medical instruments are unable to measure a biological factor below a certain known value. In the motivating Genetic and Inflammatory Markers of Sepsis (GenIMS) study, it is of interest to model the survival times of patients with community acquired pneumonia using several biomarkers and demographic covariates. In this study, the survival times are subject to right censoring while three important biomarkers are left-censored below known DLs. Our main objective in this paper is to develop a computationally-efficient procedure for conducting inference in the context of survival models with covariates subject to DLs.

© 2013 Elsevier B.V. All rights reserved.

*Corresponding author. Tel: 610 519 7343; Fax: 610 519 6928 paul.bernhardt@villanova.edu.

¹Present address: Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA

Supplementary Material The reader is referred to the online Supplementary Material for technical appendices and to <http://www.homepage.villanova.edu/paul.bernhardt/Resources2.htm> for simulation code, application code, and data similar to that analyzed in the application.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

URL: <http://www.homepage.villanova.edu/paul.bernhardt> (Paul W. Bernhardt)

Traditionally, a few common approaches have been taken to handle censoring due to DLs. Mostly simply, a complete case approach can be applied, where data from individuals with censored covariates are completely discarded. While we prove in Section 3.1 that a complete case analysis leads to consistent estimates in accelerated failure time models when covariates are censored due to DLs, efficiency is lost due to the deletion of data. In an attempt to use all the data, a second approach has become increasingly common where censored observations are replaced with a fixed value such as the DL, $DL/2$, or $DL/\sqrt{2}$ (Hornung and Reed, 1990) or the conditional mean of the censored covariate (Austin and Hoch, 2004; Giovanini, 2008; Arunajadai and Rauh, 2012). Using these naive substitution methods has been demonstrated to produce biased parameter estimates in generalized linear models (Lynn, 2001; Austin and Brunner, 2003; Lubin et al., 2004; Rigobon and Stoker, 2007, 2009; Helsel, 2012) and survival models (D'Angelo and Weissfeld, 2008; Sattar et al., 2012).

A few researchers have considered alternative methods for handling censored covariates in survival models. Langohr et al. (2004) and Sattar et al. (2012) proposed fully-parametric survival models with a single interval-censored predictor. Lee et al. (2003) also considered survival models with a single covariate subject to DLs, though they proposed using semiparametric Cox proportional hazards models in which the relative risk function for the censored covariates is replaced by a nonparametric estimate of its expected value. More recently, D'Angelo and Weissfeld (2008) developed an indexing approach where censored covariate values are directly replaced by their conditional expectation given a linear combination of the fully observed covariates. While their method performs reasonably well, it is somewhat ad-hoc and limited to cases when no more than two covariates are subject to DLs.

In this paper, we develop a straightforward, computationally-efficient multiple imputation method for handling multiple covariates subject to DLs in the context of accelerated failure time (AFT) models for censored survival data. To increase flexibility in the AFT survival model, we recommend using the seminonparametric (SNP) distribution to model the error term. We establish the asymptotic consistency and normality of the multiple imputation estimator and propose a convenient variance estimation method. We additionally suggest an iterative version of this estimator which improves efficiency with only a few updates. Though multiple imputation has been studied in many missing-data problems, our development for censored covariates in flexible survival models is nonstandard. Through numerical studies, we demonstrate that our proposed estimator leads to unbiased estimates and is potentially more efficient than several competing methods. We additionally show that using the flexible SNP distribution is more robust than typical parametric methods.

The remainder of the paper is organized as follows. In Section 2, we review AFT models and the SNP distribution. We also briefly explain how to fit AFT models with an SNP error term. In Section 3, we develop the proposed multiple imputation methods and establish their asymptotic properties. In Section 4, we carry out extensive simulations to compare the performance of the proposed methods with several simpler approaches. In Section 5, we apply the proposed methods to the dataset from the GenIMS study. Finally, in Section 6, we discuss the limitations of the proposed multiple imputation methods and some avenues for further research. The technical details for the proposition and theorems appearing in this paper are provided in the online Supplementary Material.

2. Seminonparametric Accelerated Failure Time Model

We first present the proposed seminonparametric accelerated failure time (SNP-AFT) model and discuss the algorithm for fitting the model when covariates are fully observed.

2.1. SNP-AFT Model

The accelerated failure time (AFT) model provides a useful way of relating a q -dimensional vector of covariates, W_i to a survival time, T_i , $i = 1, \dots, n$. Specifically, the AFT model assumes that

$$\log T_i = W_i^T \beta + \sigma \epsilon_i, \quad (1)$$

where β is a vector of coefficient parameters relating W_i to T_i , σ is a scale parameter, and ϵ_i , $i = 1, \dots, n$, are independent and identically distributed errors. Due to right censoring of the survival times, we only observe $Y_i = \min(T_i, C_i)$, and $\delta_i = I(T_i \leq C_i)$, where C_i is a censoring random variable.

Traditionally, fitting an AFT model has required a fully-parametric approach, and thus semiparametric Cox proportional hazards models are often preferred in practice. However, AFT models have several advantages over Cox proportional hazards models. Perhaps most significantly, AFT models allow us to directly interpret the parameters β as representing the conditional effect of the covariates on survival time. Additionally, with AFT models we can easily derive the distribution of T_i , which is useful for accommodating covariates subject to DLs in the multiple imputation algorithm described in Section 3.2.

While we suggest using AFT models, we still wish to flexibly model the survival time T_i . To do this, we propose using the seminonparametric (SNP) distribution to flexibly model the error term ϵ_i in (1). A univariate random variable V is distributed SNP if its density function can be expressed as

$$f_K(v) = P_K^2(v) g(v), \quad (2)$$

where $P_K(v) = a_0 + a_1 v + \dots + a_K v^K$ is a polynomial of order K , $g(v)$ is a kernel density with a moment generating function, and $\int f_K(v) dv = 1$. Appropriate choices for $g(v)$ will be discussed in Section 2.2. When $K = \infty$, Gallant and Nychka (1987) showed that any “smooth” density may be represented by an SNP distribution, where smooth is defined by certain smoothness and differentiability conditions. Even when considering a finite value of K , the SNP can still effectively model a wide variety of smooth distributions including skewed, heavy- or light-tailed, or multimodal distributions.

The degree of the polynomial in (2), K , can be thought of as a tuning parameter. For larger values of K , the SNP distribution is more flexible but also more complicated. When $K = 0$, the SNP distribution is identical to that of the kernel density $g(v)$. In this paper, we suggest choosing K using the Bayesian information criterion (BIC) developed by Schwarz (1978). We recommend using the BIC because it favors simpler densities compared to most other well-known information criterion, and the computational requirements for fitting an SNP distribution are greatly reduced with smaller values of K . It has been our practice to only consider $K \geq 2$ since it was shown by Zhang and Davidian (2001) and Tang (2008) that $K > 2$ is usually not chosen by selection criterion for most distributional shapes.

While the SNP density is flexible, it can be challenging to maximize the likelihood given the observed data. Zhang and Davidian (2001) suggested an alternative parameterization to make maximization more stable. They first noted that forcing $\int f_K(v) dv = 1$ is equivalent to

requiring $E \left\{ P_K^2(U) \right\} = 1$, where $U \sim g(u)$. It can then be shown that

$E \left\{ P_K^2(U) \right\} = c^T c = c_1^2 + c_2^2 + \dots + c_K^2 = 1$ and there exists a unique transformation from c to $a = (a_0, a_1, \dots, a_K)^T$. Then, rather than estimate the $K + 1$ parameters in a subject to the aforementioned constraints, we can use a polar coordinate transformation to estimate the K

parameters in $\mathbf{c}(\varphi)$ without any constraints, where $c_1 = \sin(\varphi_1)$, $c_2 = \cos(\varphi_1) \sin(\varphi_2)$, ..., $c_K = \cos(\varphi_1) \cos(\varphi_2) \dots \sin(\varphi_K)$, and $\varphi_i \in (-\pi/2, \pi/2]$, $i = 1, \dots, K$. We refer to Zhang and Davidian (2001) for more details on deriving this alternative parameterization.

Ultimately, we want to use the SNP to model the distribution of the error term in the AFT model (1). In the next subsection, we explain how to fit an AFT model with SNP error for a given K when the covariate vector $\mathbf{W}_i$ is completely observed without censoring.

2.2. Fitting an SNP-AFT Model

To fit an SNP distribution for the error in (1), it is necessary to choose a kernel for the SNP, $g(v)$. Since when $K = 0$ the SNP distribution is identical to that of the kernel density $g(v)$, it is logical to let $g(v)$ be a standard survival distribution. In this paper, we follow the strategy of Zhang and Davidian (2007) and Doehler and Davidian (2008) by considering two possibilities for $g(v)$, the standard normal and standard exponential densities. These two distributional choices require slightly different approaches for fitting an SNP-AFT model. Specifically, the error term ϵ_i in (1) can be modeled directly using a normal-based SNP, but for the exponential-based SNP we must model $\exp(-\epsilon_i)$, since the exponential distribution only takes positive values. Then, for either modeling choice, making the usual assumption that $T_i \perp\!\!\!\perp C_i$ given $\mathbf{W}_i$, we can estimate the parameter vector $\boldsymbol{\theta} = (\beta^T, \varphi^T)^T$ by maximizing the likelihood

$$\mathcal{L}(\boldsymbol{\theta}; y_i, \delta_i, w_i) = \prod_{i=1}^n \{f_T(t_i | w_i)\}^{\delta_i} \{S_T(c_i | w_i)\}^{1-\delta_i}, \quad (3)$$

where f_T and S_T are the conditional density and survival functions of T_i respectively. Using standard transformation techniques, it is easy to show that for the normal-based SNP,

$f_T(t|w) = (\sigma t)^{-1} P_K^2(Q) g(Q)$ and $S_T(c|w) = \int_L^\infty P_K^2(z) g(z) dz$, where $Q = \{\log(t) - \mathbf{w}^T \boldsymbol{\beta}\} / \sigma$, g is a standard normal density, and $L = \{\log(c) - \mathbf{w}^T \boldsymbol{\beta}\} / \sigma$. Similarly, for the exponential-based SNP, we find that $f_T(t|w) = (\sigma t)^{-1} Q P_K^2(Q) g(Q)$ and $S_T(c|w) = \int_L^\infty P_K^2(z) g(z) dz$, where $Q = \exp[\{\log(t) - \mathbf{w}^T \boldsymbol{\beta}\} / \sigma]$, g is the standard exponential density, and $L = \exp[\{\log(c) - \mathbf{w}^T \boldsymbol{\beta}\} / \sigma]$.

Maximizing the likelihood (3) when $K \geq 1$ is not easy, largely because the SNP distribution is so flexible and the vector of parameters, $\boldsymbol{\theta}$, is potentially high-dimensional. For this reason, we suggest obtaining several sets of good starting values for $\boldsymbol{\theta} = (\beta^T, \varphi^T)^T$ and then conducting maximization algorithms from each of these sets of values. We recommend the following steps to obtain these starting values:

1. Fit the AFT model (1) with normal errors to get the initial estimates β^* and φ^* ;
2. Using β^* and φ^* , evaluate the log of the likelihood given in (3) at a grid of points in $(-\pi/2, \pi/2]^K$, a K -dimensional hypercube covering the range of values for φ ;
3. Using φ^* from the r highest log-likelihood values in step 2, update the estimates for β and the intercept β_0 by equating the moments of the normal-AFT model fit in Step 1 with the SNP-AFT model:
 - a. if the SNP is normal-based, re-estimate β as $\beta^{**} = \beta^* / \text{SD}(V)$ and β_0 as $\beta_0^{**} = \beta_0^* - \sigma^{**} E(V)$, where $E(V)$ and $\text{SD}(V)$ are the mean and standard deviation of a variable V distributed as a normal-based SNP;
 - b. if the SNP is exponential-based, re-estimate β as $\beta^{**} = \beta^* / \text{SD}\{\log(V)\}$ and β_0 as $\beta_0^{**} = \beta_0^* - \sigma^{**} E\{\log(V)\}$, where V is distributed as an exponential-based SNP.

Using these r sets of starting values, $(\beta_0^{**}, \beta_1^*, \dots, \beta_p^*, \sigma^{**}, \phi^*)$, standard maximization procedures can be used to find r sets of potential maximum likelihood estimates. The estimates that give the highest log-likelihood are then considered to be the actual maximum likelihood estimates. We recommend letting the grid search in Step 2 be fairly dense, with at least 10^K points, and selecting r values in Step 3 that are not “close” to each other, where close is defined by the distance between points on the grid. This prevents the procedure described above from using r sets of starting values that are relatively similar. Based on our personal experience, $3 \leq r \leq 5$ provides a good choice. This procedure should then be repeated for each value of K and for both the exponential and normal kernels. As we suggested in Section 2.1, BIC can be used to choose an appropriate kernel and value for K .

While this maximization procedure appears complicated, it can be automated quite easily and is reasonably fast. We note that it is not necessary to use numerical integration techniques to obtain $S_T(c|w)$ since there exist closed form representations for both the normal- and exponential-based SNP. Specifically, for the normal-based SNP with $K = 2$,

$$S_T(c|w) = a_0^2 \Psi(L) + 2a_0 a_1 g(L) + (a_1^2 + 2a_0 a_2) \{L g(L) + \Psi(L)\} + 2a_1 a_2 (L^2 + 2) g(L) + a_2^2 \{L^3 g(L) + 3L g(L) + 3\Psi(L)\},$$

where Ψ is the survival function for the standard normal distribution, g is the standard normal density, and L is defined as before. Similarly, for the exponential-based SNP with $K = 2$,

$$S_T(c|w) = a_0^2 \Psi(L; 1) + 2a_0 a_1 \Psi(L; 2) + 2(a_1^2 + 2a_0 a_2) \Psi(L; 3) + 12a_1 a_2 \Psi(L; 4) + 24a_2^2 \Psi(L; 5),$$

where $\Psi(\cdot; j)$ is the survival function of a gamma density with scale parameter 1 and shape parameter j and L is defined as before. To obtain $S_T(c|w)$ for $K = 1$, we can let simply $a_2 = 0$ in the expressions above.

3. Proposed Methods

3.1. Problem Set-up

For the remainder of this paper, we assume that some covariates in W_i are subject to censoring due to lower DLs. Thus, for the i^{th} individual, we let $W_i = (Z_i^T, X_i^T)^T$, where $Z_i = (Z_{i1}, \dots, Z_{i(p-q)})^T$ is the $(p-q)$ -dimensional vector of covariates fully observed for each individual and $X_i = (X_{i1}, \dots, X_{iq})^T$ is the q -dimensional vector of covariates subject to censoring below $d = (d_1, \dots, d_p)^T$, the vector of DLs. For X_{ij} , $j = 1, \dots, q$, we only observe $X_{ij}^* = \max(X_{ij}, d_j)$ and $\delta_{ij} = I(X_{ij} = d_j)$, so that the complete set of observed data for the i^{th} individual is $(Y_i, Z_i, X_i^*, \delta_i)$, where $X_i^* = (X_{i1}^*, \dots, X_{iq}^*)^T$ and $\delta_i = (\delta_{i1}, \dots, \delta_{iq})^T$. For the i^{th} subject, we define X_i^c as the subset of X_i for which $\delta_{ij} = 0$ and X_i^o as the subset of X_i for which $\delta_{ij} = 1$. Similarly, we let d_i^c be the subset of d corresponding to X_i^c .

For notational simplicity, we assume that the DLs do not vary across subjects. However, the methods discussed in this paper can be easily applied to data with subject-specific DLs as long as the DLs are known for each subject. Additionally, though we focus on lower DLs, the methods we suggest can be generalized to covariates censored due to upper DLs. Finally, in this paper we assume that the distribution $f(x|z, \theta)$ is a known q -variate distribution

indexed by the parameter vector γ , from which we can obtain $f(x_i^c|z_i, x_i^0; \gamma)$ for all possible subsets of censored covariates, X_i^c . This distributional assumption provides a logical way to use information in the observed data region above the DLs to extrapolate below the DLs. Often it is appropriate with biological variables to assume that $f(x_i|z_i)$ is multivariate normal after a Box-Cox transformation. However, in situations where assumption is inappropriate, we recommend using a series of conditional univariate distributions to represent $f(x_i|z_i)$ as $f(x_{i1}|x_{i2}, x_{i3}, \dots, x_{iq}, z_i) f(x_{i2}|x_{i3}, \dots, x_{iq}, z_i) \dots f(x_{iq}|z_i)$. Then, each of these conditional distributions can be modeled by a regression model with flexible errors. We henceforth denote the complete set of parameters by $\gamma = (T, \phi^T, T)T$.

We remind the reader that our main objective in this paper is to estimate γ , the regression parameters in the AFT model (1). To estimate these parameters, we have suggested using the SNP distribution to describe the random error in the model. In the following, we briefly review maximum likelihood and complete case analyses when a subset of the covariates, X_i , is subject to DLs, and we explain why these two approaches are not ideal for handling covariates censored due to DLs in survival analyses.

The likelihood contributed by all the observed data can be represented as

$$\prod_{i=1}^n \int_{-\infty}^{d_i^c} \left\{ f_T(t_i|z_i, x_i^0, x_i^c; \beta, \sigma, \phi) \right\}^{\delta_i} \left\{ S_T(c_i|z_i, x_i^0, x_i^c; \beta, \sigma, \phi) \right\}^{1-\delta_i} f(x_i^0, x_i^c|z_i; \gamma) dx_i^c, \quad (4)$$

where f_T and S_T are the conditional density and survival functions of T_i and $\int_{-\infty}^{d_i^c}$ is an integral whose dimension corresponds to the length of X_i^c . The likelihood (4) potentially involves high-dimensional integration and generally does not have a closed form. For this reason we often must resort to numerical integration when maximizing (4), which can be extremely slow.

As an alternative to maximizing the likelihood using the data, the simplest way to handle data subject to DLs is to consider only the complete cases. In the context of survival data, the complete cases are defined with respect to the covariates, not the survival variable. Thus, the complete case estimator, $\hat{\beta}_{CC}$, is found by maximizing the likelihood

$$\prod_{i=1}^n \rho_i^* \left\{ f_T(t_i|z_i, x_i) \right\}^{\delta_i} \left\{ S_T(c_i|z_i, x_i) \right\}^{1-\delta_i}, \quad (5)$$

where $\rho_i^* = \prod_{j=1}^q \rho_{ij}$ and $\rho_{ij} = I(X_{ij} > d_j)$.

Proposition 1—Suppose assumptions B.1-B.3 given in the online Supplementary Material hold. Then, for an AFT model (1) with a correctly specified error distribution, $\hat{\beta}_{CC} \xrightarrow{P} \beta$ as $n^* \rightarrow \infty$, where β is the true parameter vector and n^* is the complete case sample size. $\parallel$ Since the censoring on X_i does not depend on the survival time T_i , the complete case estimator is still consistent. However, if a large percentage of individuals have censored covariates, the complete case estimator may have substantial efficiency loss.

In the next subsection, we present our proposed multiple imputation algorithm for handling covariates subject to DLs in AFT survival models. This method reduces the computational burden compared to the maximum likelihood method and increases the efficiency compared to the complete case method.

3.2. Multiple Imputation Method

Our proposed multiple imputation method is rooted in the fact that the complete case estimator for the parameters in the AFT model is consistent. This allows us to fairly easily obtain a consistent set of initial estimates for $\theta = (T, \phi^T, T)^T, \hat{\theta}^I$, which can then be used to generate m sets of imputed data for the censored covariates from the distribution

$f(x_i^c | y_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}^I)$. Using the m sets of imputed data for the censored covariates, $X_i^c, i=1, \dots, n$, final estimates for θ can be obtained using standard techniques.

Our proposed multiple imputation algorithm is as follows:

Step 1—Using all the data $(z_i, x_i^*, \rho_i), i = 1, \dots, n$, and the DLs d_i , obtain the maximum likelihood estimate, $\hat{\gamma}^I$, as an initial estimate for γ , the parameters indexing $f(x_i | z_i)$.

Step 2—Using the complete cases, maximize the likelihood given by (3) using the SNP distribution for $f(y_i, z_i, x_i)$ with $K = 0, 1, 2$ for both the normal and exponential kernels. Obtain the estimates, $\hat{\beta}^*, \hat{\sigma}^*$, and $\hat{\phi}^*$, for each kernel and K . Set $\hat{\beta}^I$, and $\hat{\sigma}^I$ equal to the estimates that minimize the BIC across K and the two different kernel models.

Step 3—Form the 1st imputed dataset by generating $\tilde{x}_i^c$ for each individual i with censoring, $i = 1, \dots, n$, from $f(x_i^c | y_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}^I)$ to replace the censored covariates x_i^c where $\hat{\theta}^I = (\hat{\beta}^{IT}, \hat{\sigma}^I, \hat{\phi}^{IT}, \hat{\gamma}^{IT})^T$. Denote $\tilde{x}_i$ as the resulting covariate vector for the i^{th} subject that include both x_i^o and $\tilde{x}_i^c$.

Step 4—Using the complete dataset, $(y_i, z_i, \tilde{x}_i), i = 1, \dots, n$, estimate $\theta = (T, \phi^T, T)^T$ as $\hat{\theta}^{(1)}$ by maximizing the likelihood (3).

Step 5—Repeat Steps 3 and 4 m times, resulting in $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(m)}$.

Step 6—Finally, the multiple imputation estimator is defined as $\hat{\theta}_{MI} = m^{-1} \sum_{k=1}^m \hat{\theta}^{(k)}$.

Steps 2-4 of this algorithm warrant additional explanation. In Step 2, the maximum likelihood estimates for the complete cases can be obtained exactly as explained in Section 2.2. The algorithm for finding starting values needs to be employed for both the normal and exponential kernel as well as $K = 0, 1, 2$. In Step 4, while K and the kernel for the SNP have already been fixed, several sets of starting values still need to be used to ensure that the maximum likelihood estimates are found. These starting values can be found similarly as in the algorithm described in Section 2.2, except that the SNP-AFT model estimated by the complete cases should be used for obtaining initial estimates rather than the normal-AFT model.

In Step 3, we can obtain samples from $f(x_i^c | y_i, \delta_i, z_i, x_i^o, x_i^c < d_i^c; \hat{\theta}^I)$ using standard Monte Carlo sampling methods such as the Metropolis-Hastings algorithm. However, we must be careful in how we do this. Specifically, in order to avoid modeling the distribution of the censoring variable C_i we need to make the following assumptions:

A.1 $T_i \perp\!\!\!\perp C_i$ given W_i

A.2 $C_i \perp\!\!\!\perp X_i$ given Z_i

A.3 the conditional distributions for T_i and C_i given W_i have no common parameters, where $\perp\!\!\!\perp$ indicates statistical independence.

Then, if $\delta_i = 1$ ($T_i < C_i$),

$$f(x_i^c | t_i, \delta_i=1, z_i, x_i^o) = \frac{f_T(t_i | z_i, x_i) S_C(t_i | z_i, x_i^o, x_i^c) f(x_i^c | z_i, x_i^o)}{f(t_i, \delta_i=1 | z_i, x_i^o)} \propto f_T(t_i | z_i, x_i) f(x_i^c | z_i, x_i^o),$$

and if $\delta_i = 0$ ($T_i > C_i$),

$$f(x_i^c | c_i, \delta_i=0, z_i, x_i^o) = \frac{S_T(c_i | z_i, x_i) f_C(c_i | z_i, x_i^o, x_i^c) f(x_i^c | z_i, x_i^o)}{f(c_i, \delta_i=0 | z_i, x_i^o)} \propto S_T(c_i | z_i, x_i) f(x_i^c | z_i, x_i^o),$$

While A.1 and A.3 are common assumptions in survival analyses and allow us to avoid modeling the censoring variable C_i when the covariates are fully observed, we now also need A.2 so that f_C and S_C no longer involve X_i^c . Note that these same assumptions are also necessary for the maximum likelihood method described in Section 3.1.

In many cases, it may be reasonable to believe that X_i is not related to C_i . However, even if they are somehow dependent, we can hope that the fully observed covariates Z_i completely explains the extent of this relationship. If A.2 is questionable, we suggest fitting a flexible parametric model for C_i so that the assumption is no longer necessary.

For multiple imputation algorithms conducted using Bayesian methods, it is typical to use the standard error estimation formula proposed by Rubin (1987),

$$\widehat{\text{Var}}(\hat{\theta}) = V_W + (1+m^{-1}) V_B, \quad (6)$$

where V_W , the within imputation variance, is the average of the estimated covariance matrices for $\hat{\theta}^{(k)}$, $k = 1, \dots, m$, and V_B , the between imputation variance, is the sample variance-covariance matrix of $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(m)}$. However, Rubin's variance formula (6) is not a consistent estimator for the variance of our proposed estimator $\hat{\theta}_{MI}$ because it does not take into account the variation in the initial estimates used in Step 1 of the procedure.

The following theorem provides the asymptotic properties of $\hat{\theta}_{MI}$ and allows us to construct a consistent variance estimator for our multiple imputation algorithm.

Theorem 1—Suppose assumptions A.1-A.3 and B.1-B.3 given in the online Supplementary Material hold. Then, for an AFT model (1) with a correctly specified error distribution,

$$n^{1/2} (\hat{\theta}_{MI} - \theta) \xrightarrow{d} N(0, \Sigma),$$

where

$$\Sigma = \{I^F(\theta)\}^{-1} + (1+m^{-1}) \{I^F(\theta)\}^{-1} D\{I^F(\theta)\}^{-1} + \{I^F(\theta)\}^{-1} D\text{Var}(\hat{\theta}^I) D\{I^F(\theta)\}^{-1}, \quad (7)$$

$D = \{I_F(\cdot) - I(\cdot)\}$, $\hat{F}(\cdot) = E\{S^F(y, \mathbf{w}, \cdot)S^{FT}(y, \mathbf{w}, \cdot)\}$ is the full-data information matrix derived from the likelihood without censoring on the covariates, and $I(\cdot) = E\{S(y, \mathbf{z}, \mathbf{x}^*, \cdot)S^T(y, \mathbf{z}, \mathbf{x}^*, \cdot)\}$ is the observed-data information matrix derived from the likelihood with censoring on the covariates, evaluated at the true parameter vector, $\cdot$. Here S^F and S denote the full- and observed-data score functions.

While we note that we rarely can *a priori* know the true error term for the AFT model (1), using the SNP distribution as described in Section 2 allows us to well approximate the error distribution and derive approximately consistent estimates for the parameters. To estimate (7), we can use consistent estimators for $\hat{F}(\cdot)$ and $I(\cdot)$ similar to those proposed by Wang and Robins (1998) and Robins and Wang (2000) in the context of missing data,

$$\hat{I}^F(\theta_0) = \left[-m^{-1} \sum_{k=1}^m \left\{ n^{-1} \sum_{i=1}^n \frac{\partial S^F(y_i, \delta_i, z_i, \tilde{x}_{ik}, \hat{\theta}^{(k)})}{\partial \theta^T} \right\}^{-1} \right]^{-1}, \quad (8)$$

$$\hat{I}(\theta_0) = n^{-1} \sum_{i=1}^n \{m(m-1)\}^{-1} \sum_{k, k'=1, \dots, m, k \neq k'} S^F(y_i, \delta_i, z_i, \tilde{x}_{ik}, \hat{\theta}^{(k)}) S^{FT}(y_i, \delta_i, z_i, \tilde{x}_{ik'}, \hat{\theta}^{(k')}),$$

where $S^F = \frac{\partial}{\partial \theta} \{\ell(\theta; y_i, \delta_i, z_i, \tilde{x}_{ik})\}$ is the score vector from the full data with log-likelihood $\ell(\theta; y_i, \delta_i, z_i, \tilde{x}_{ik})$, $\tilde{x}_{ik}$ denotes the k^{th} vector of imputed values for the i^{th} individual, and $\hat{\theta}^{(k)}$

denotes the estimated parameter vector based on the k^{th} dataset. Also, $(\hat{\theta}^I)$ can be estimated by the observed information matrix for the initial parameter estimates $\hat{\theta}^I$. It may seem counterintuitive to estimate $\hat{F}(\cdot)$ by the inverse of the average of the inverse information matrices for the m imputed datasets rather than the average of the information matrices. However, the suggested estimator is more stable when using the SNP distribution because the estimated φ parameters may possibly vary widely across imputed datasets. By taking the inverse before averaging, we are able to maintain very good final estimates for the sub-information matrix for the φ parameters, which are of primary interest.

Since it is often laborious to estimate the asymptotic variance using (8), we propose an alternative easy-to-calculate approximate variance formula,

$$\hat{\Sigma} = V_W + (1+m^{-1}) V_B + V_B V_W^{-1} \text{Var}(\hat{\theta}^I) V_W^{-1} V_B, \quad (9)$$

where V_W and V_B are the within- and between-imputation variances, respectively. This estimator can be derived by noting that Rubin's variance formula (6) is consistent for the first two terms of (7) as $n \rightarrow \infty$ and $m \rightarrow \infty$ (Wang and Robins 1998). Though (9) is not a consistent estimator finite m , we have found that it works well even for a high percentage of censoring on the covariates ($> 50\%$) and a moderate number of imputed data sets, say $m \in [15, 30]$.

3.3. Iterated Multiple Imputation Method

As $m \rightarrow \infty$, the multiple imputation estimator described in the previous section is equivalent to a one-step update of the EM algorithm (Wang and Robins, 1998; Tsiatis, 2006). Thus, we can iterate the multiple imputation procedure with a reasonably large m to approximate maximum likelihood estimation of the model parameters, $\cdot$. In practice, we may wish to only perform a few iterations, with a relatively small m at each iteration (say $m = 50$), in order to balance computational time with the efficiency gained by iterating the multiple imputation procedure.

In the following iterated multiple imputation procedure, define an s -dimensional vector m , where s corresponds to the number of updates for the procedure and m_i is the number of imputed datasets used for each update, $i = 1, \dots, s$. To guarantee increased efficiency with each update, we let $m_1 \leq m_2 \leq \dots \leq m_s$.

Step 1—To find the 1st update for θ , $\hat{\theta}^{(1)}$, conduct the complete multiple imputation procedure described in Section 3.2 with m_1 imputed datasets. Obtain the associated variance-covariance matrix, $\widehat{\text{Var}}(\hat{\theta}^{(1)})$.

Step 2—To find the 2nd update for θ , $\hat{\theta}^{(2)}$, repeat the multiple imputation procedure with m_2 imputed datasets, but use $\hat{\theta}^{(1)}$ and $\widehat{\text{Var}}(\hat{\theta}^{(1)})$ as the initial parameter and variance estimates.

Step 3—To find the i^{th} update for θ , $\hat{\theta}^{(i)}$, repeat the multiple imputation procedure with m_i imputed datasets, using $\hat{\theta}^{(i-1)}$ and $\widehat{\text{Var}}(\hat{\theta}^{(i-1)})$ as the initial parameter and variance estimates.

Step 4—Let $\hat{\theta}_{IMI} = \hat{\theta}^{(s)}$ and $\widehat{\text{Var}}(\hat{\theta}_{IMI}) = \widehat{\text{Var}}(\hat{\theta}^{(s)})$ be the final iterated multiple imputation parameter and variance and estimates.

In the procedure above, the final variance estimator is easy to calculate since the estimated variance for the parameter estimates at each step is simply an update of the estimate from the previous step. Either (8) or (9) can be used to estimate the variance of the updated multiple imputation estimator at each step. The following result gives theoretical justification for using the iterated multiple imputation estimator.

Theorem 2—Suppose assumptions A.1-A.3 and B.1-B.3 given in the online Supplementary Material hold. Then, for an AFT model (1) with a correctly specified error distribution,

$$n^{1/2}(\hat{\theta}_{IMI} - \theta) \xrightarrow{d} N(0, \Sigma_s),$$

where

$$\Sigma_s = \{I^F(\theta)\}^{-1} + (1+m^{-1})\{I^F(\theta)\}^{-1}D\{I^F(\theta)\}^{-1} + \{I^F(\theta)\}^{-1}D\widehat{\text{Var}}(\hat{\theta}^{(s-1)})D\{I^F(\theta)\}^{-1}, \quad (10)$$

$\widehat{\text{Var}}(\hat{\theta}^{(s-1)})$ is the variance of $\hat{\theta}^{(s-1)}$, the estimator for θ at the $(s-1)^{\text{th}}$ iteration, and all other quantities are defined as in Theorem 1.

While the iterated multiple imputation procedure and associated theory described above are based on a known number of iterations s , it is not required for s to be predetermined. That is, the same procedure and theory applies if instead choosing s before conducting the analysis, we continue iterating the multiple imputation procedure until the difference between consecutive estimates is negligible, such as when

$$\max \left(\frac{|\theta_1^{(s)} - \theta_1^{(s-1)}|}{|\theta_1^{(s-1)}|}, \dots, \frac{|\theta_t^{(s)} - \theta_t^{(s-1)}|}{|\theta_t^{(s-1)}|} \right) < \epsilon,$$

where t is the total number of parameters and ϵ is a predefined distance. In this case, as described previously, the iterated multiple imputation estimator will simply produce the maximum likelihood estimates since it is equivalent to an EM algorithm. However, the computational requirements for obtaining the maximum likelihood estimates via this iterated multiple imputation procedure may be burdensome if s is large, especially since the number of iterations m at each step also needs to be large in order to guarantee convergence. For this reason, we recommend pre-choosing s to be relatively small unless computational time is not an issue.

4. Simulation

We conducted numerical studies to assess the performance of the proposed multiple imputation and iterated multiple imputation estimators. We set up the simulations to represent a situation similar to that in the application described in Section 5. Specifically, we generated the covariates $Z \sim \text{beta}(3, 2) \cdot 84 + 18$, to represent an “age” variable, and

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left\{ \begin{pmatrix} 1.4 + 0.0087Z \\ 2.4 - 0.0091Z \end{pmatrix}, \begin{pmatrix} 2.5 & 1 \\ 1 & 5 \end{pmatrix} \right\}$$

to represent the log-cytokines TNF and IL-10. Observations of X_1 and X_2 were censored at the DLs $d_1 = \log(4)$ and $d_2 = \log(5)$, with censoring rates of 35% and 47%. The distributions for these covariates, as well as the DLs, censoring rates, and parameter values, were chosen specifically to resemble the GenIMS data. Finally, we generated the survival times based on the model $\log(T) = (1, X_1, X_2, Z) \beta + \epsilon$, where $\beta = (5, -0.2, 0.2, -0.1)^T$. For the first simulation design, we generated $\epsilon \sim \text{Weibull}(1.5)$, where ϵ is the scale parameter in the AFT model (1). For the second simulation design, we generated $\epsilon \sim p\text{Weibull}(2) + (1-p)\text{lognormal}(0, 1)$, a mixture distribution with $p \sim \text{Bernoulli}(0.6)$. For both simulation designs, we generated a censoring variable C from a uniform distribution so that approximately 60% of the survival times were censored. We simulated 1000 datasets with 500 individuals per dataset.

We compared several estimation methods: (i) complete case method (CC); (ii) substitution method (SUB) using DL/2 to replace the censored covariate values; (iii) multiple imputation method (MI); (iv) iterated multiple imputation method (IMI). For the MI estimator, we obtained $m = 15$ imputed datasets while for the IMI estimator, we performed $s = 4$ iterations with $m = (15, 15, 30, 50)^T$. We chose $s = 4$ to illustrate the potential efficiency gains by iterating the multiple imputation procedure while still remaining computationally reasonable.

For each of the methods, we considered three strategies for obtaining final estimates. First, we analyzed the survival data assuming a Weibull distribution for the error in the AFT model. Second, we considered fitting the AFT model with Exponential, Weibull, lognormal, and loglogistic distributed error terms and choosing between these distributions using the BIC. We label this strategy the “parametric” method. Finally, we analyzed the survival data using the SNP distribution for the error term. We estimated the model parameters using the CC, SUB, MI, and IMI methods for each of these three strategies. For the CC method, we

maximized the likelihood given by (5) while for the SUB, MI, and IMI methods we maximized the likelihood (3) after obtaining imputations.

Table 1 gives the average bias, Monte Carlo standard deviation (SD), and average standard error (SE) estimates for the 1000 simulated datasets for both the Weibull and mixture errors. These estimates are given for the CC, SUB, MI, and IMI analysis methods based on the Weibull, parametric, and SNP distributional assumptions described above.

For the simulation design with Weibull errors, Table 1 shows that the CC analysis leads to very inefficient estimates while the SUB method leads to biased parameter estimates. The parameter estimates are unbiased for the CC, MI, and IMI methods for all three modeling strategies. While correctly specifying the error distribution as Weibull gives lower standard errors, the parametric and SNP fitting strategies perform competitively. For the non-intercept parameters of the AFT model, using the Weibull only offers 1-6% relative efficiency gains over the SNP, with even less improvement over the parametric method,

where the relative efficiency gain is calculated by $\left\{ \text{var}(\hat{\beta}_{\text{SNP}}) - \text{var}(\hat{\beta}_{\text{Weib}}) \right\} / \text{var}(\hat{\beta}_{\text{SNP}})$. Regardless of the strategy that is used for fitting the error term, the IMI methods offers substantial efficiency gains over both the MI and CC analysis methods. For example, with an SNP analysis the relative efficiency gains are 73-79% and 9-30% over the CC and MI methods, respectively. Additionally, note that the efficiency of the IMI method for the SNP strategy is nearly identical to correctly assuming Weibull errors in the AFT model.

For the simulation design with mixture errors, the CC method gives unbiased estimates for the parametric and SNP fitting methods, but the intercept term is biased when incorrectly assuming the Weibull model. The CC estimator is inefficient for all three fitting strategies and the standard error estimation is poor when taking the Weibull or parametric approach. As before, the SUB method gives poor parameter and standard error estimates in all scenarios. Since the Weibull distribution inaccurately models the error in the AFT model, it does poorly in terms of bias and standard error estimation even with the MI and IMI methods. Interestingly, both the parametric and SNP analysis strategies do fairly well in terms of bias for the MI and IMI methods, though the parametric approach leads to higher standard errors. Finally, we note that the IMI procedure outperforms both the CC and MI analysis in terms of efficiency. For example, with an SNP analysis the relative efficiency gains are 72-87% and 14-32% over the CC and MI methods, respectively.

Overall, it is clear that the multiple imputation methods provide significant efficiency gains over the complete case method. Additionally, fitting the SNP-AFT model provides bias and efficiency improvements over more restrictive parametric approaches when the survival times do not follow a usual parametric form.

The simulation above was implemented in R on a computer with an Intel Core i7 870 @ 2.93 GHz processor and 8 GB of RAM. While most of the R functions for implementing the procedures were self-written, the function “Metrop1R” in the package “MCMCpack” (Martin et al., 2011) is useful for obtaining imputed samples in Step 3 of the proposed multiple imputation algorithm described in Section 3.2 and the “survreg” function in the package “survival” (Therneau, 2012) is useful for fitting AFT survival models with standard distributions. For the scenario where the true survival distribution is Weibull, the average computational times per dataset were 0.19 and 0.45 minutes for the parametric and SNP MI methods, respectively, and 0.73 and 1.25 minutes for the parametric and SNP IMI methods, respectively. For the scenario where the true survival distribution is a mixture distribution, the average computational times per dataset were 0.20 and 2.10 minutes for the parametric and SNP MI methods, respectively, and 0.74 and 14.6 minutes for the parametric and SNP IMI methods, respectively. The computational times when assuming a Weibull distribution

were very similar to the parametric case, and all of the average computational times for the CC and SUB analysis methods were less than 0.2 minutes.

5. Application to GenIMS data

We demonstrate the performance of our proposed multiple imputation and iterated multiple imputation methods by applying them to the data from the Genetic and Inflammatory Markers of Sepsis (GenIMS) study. One of the purposes of the GenIMS study was to identify the relationship between the survival time of patients with community acquired pneumonia (CAP) and several biomarkers for inflammatory responses in the body (Kellum et al., 2007). The data for the GenIMS study were obtained from individuals hospitalized with CAP. Several biological measurements and basic demographic information were recorded for each individual. In our analysis, we included the 1418 patients who necessitated a hospital stay, actually acquired CAP, and had measurements for the biomarkers of interest on the first day of hospitalization.

Three biomarkers for inflammatory responses in the body were measured in the GenIMS study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). These biomarkers are specific cytokines, which are cell-signaling protein molecules used by the immune system to communicate with the rest of the body. The TNF and IL-6 cytokines are biomarkers of pro-inflammatory responses to CAP while IL-10 is a biomarker of anti-inflammatory responses to CAP.

We used an AFT regression model with an SNP error term to explain the relationship between survival time and six covariates: the levels of the three cytokines, TNF, IL-6, and IL-10, on the first day of hospitalization, sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), and age. While sex, race, and age were fully observed, TNF, IL-6, and IL-10 were all subject to censoring below the detection thresholds 4, 2 or 5, and 5 pg/ml, leading to 35.54%, 13.40%, and 46.83% censoring, respectively. Approximately 64% of the individuals had at least one of these three measurements censored.

In our analysis, we assumed that the natural logarithms of TNF, IL-6, and IL-10 are multivariate normal conditional on the sex, race, and age covariates. Though our analysis was based on the logarithms of TNF, IL-6, and IL-10, we continue to refer to the log-cytokines simply as TNF, IL-6, and IL-10. Specifically, we assumed that, given $\mathbf{z}_i$,

$(\text{TNF}_i, \text{IL-6}_i, \text{IL-10}_i)^T \stackrel{\text{ind}}{\sim} \mathbf{\mu} + \mathbf{\Sigma}^{1/2} \mathbf{z}_i$ where $\mathbf{\mu}$ is a 3×4 matrix of mean parameters, $\mathbf{z}_i = (1, \text{sex}_i, \text{race}_i, \text{age}_i)^T$, and $\mathbf{\Sigma}$ is a 3×3 covariance matrix. To informally check this normality assumption, we created normal Q-Q plots for each the cytokines, shown in Figure 1. These Q-Q plots do not include the sample quantiles below the p_0 th quantile, where p_0 is the sample censoring proportion, as the lower quantiles cannot be estimated nonparametrically. The Q-Q plots indicate that the marginal distributions of the TNF, IL-6, and IL-10 covariates are approximately normal.

Using the multivariate normal distributional assumption for the censored covariates, we conducted an AFT analysis using the proposed multiple imputation method (IMI) and iterated multiple imputation method (IMI), as well as using a complete case analysis (CC; 511 individuals). The original intent of the GenIMS study was to model survival at 90 days since pneumonia is generally a recoverable disease and deaths after 90 days are likely unrelated to the original sickness. For this reason, our analysis below is based on treating all survival times as censored if the patient is alive at 90 days. Since the censoring variable is not random, it should be obvious that it is independent of the biomarkers. Thus, the second assumption given in Section 3.2 is valid for this application.

We observe in Table 2 that all three analysis methods produce similar parameter estimates, but the standard errors for the CC method are clearly higher. The MI and IMI methods indicate that sex and age are important covariates, with males and older patients having lower survival. Specifically, the IMI analysis method gives p-values of 0.028 and less than 0.001, respectively, for these two covariates.

Individually, all three of the cytokine biomarkers, TNF, IL-6, or IL-10, are moderately significant, with p-values of 0.159, 0.091, and 0.097, respectively, based on the IMI method. The p-value for the test that all three biomarkers have no effect on survival, $H_0: \text{TNF} = \text{IL-6} = \text{IL-10} = 0$, is less than 0.001. Thus, there is significant evidence that at least one of the three biomarkers is associated with survival. Since there are moderate correlations between the cytokines, ranging from 0.22 to 0.40 in absolute value, we repeated the analysis including only one of the biomarkers in the model at a time and found that the associated p-values for TNF, IL-6, and IL-10 based on the proposed IMI method are all less than 0.001.

The analysis above was implemented in R on a computer with an Intel Core i7 870 @ 2.93 GHz processor and 8 GB of RAM. The computational times were 5.94 minutes for the complete case analysis, 8.73 minutes for the MI analysis, and 19.44 minutes for the IMI analysis.

6. Discussion

We have proposed a multiple imputation method for handling covariates censored due to DLs in AFT survival models. We have proven that the proposed estimator is consistent and asymptotically normal, with standard errors that are relatively easy to estimate. We also suggested an iterated version of the multiple imputation procedure which provides potentially significant efficiency improvements. Through numerical studies, we demonstrated that the multiple imputation procedure and iterated multiple imputation procedure provide efficiency gains over the complete case approach and that substitution methods lead to biased results.

For the simulation study in Section 4, we chose specific values for the DLs, censoring percentages, and parameters to represent a situation similar to that in the GenIMS dataset. Thus, while the simulation study demonstrates the potential advantages of the proposed multiple imputation methods in cases similar to that observed in reality for the GenIMS study, it is limited in the sense that we did not explore a wide variety of parameter and censoring scenarios. However, we did show theoretically that our proposed methods should have good asymptotic properties in a variety of cases.

We note that our proposed multiple imputation algorithm could be adapted to include imputations for censored T_i 's. However, imputing values for T_i generally requires more computation than maximizing the likelihood (3) which includes individuals for which $\delta_i = 0$. Additionally, multiple imputation methods are less efficient than maximum likelihood methods and should generally only be used when maximum likelihood methods are infeasible.

The validity of the complete case estimator in our set-up does not rely on any parametric assumption for the distribution $f(\mathbf{x}_j | z_j)$. We note that in practice it is difficult to assess the parametric distributional assumptions for observations below the DL. Therefore, if the censoring level on X_j is not high, the complete case analysis would be a good option since it is robust, valid, and easy to apply. However, in cases with heavy censoring on X_j , the proposed multiple imputation methods are preferred to improve the efficiency.

Lastly, we note that if in addition to the censored covariates there is also data that is missing at random, the complete case estimator and the multiple imputation methods may no longer be appropriate if missingness depends on the response. Fortunately, the iterated multiple imputation estimator does not require a consistent initial estimate since, for a large number of iterations and a large number of imputations at each iteration, it is equivalent to an EM algorithm. However, when the initial estimator is not consistent a different formula for variance estimation than that suggested in Section 3.2 must be used, such as that suggested by Louis (1982).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to express their appreciation to the editor, an associate editor, and two anonymous referees for their valuable comments. The authors would also like to thank Dr. Lan Kong of Penn State College of Medicine and the CRISMA (Clinical Research, Investigation, and Systems Modeling of Acute Illness) Center at the University of Pittsburgh for providing us with the GenIMS dataset. The research of Wang is supported by the NSF Award DMS-1007420 and NSF Career Award DMS-1149355. The research of Zhang is supported by the grants NIH R01 CA85848-12 and NIH/NIAID R37 AI031789-20.

References

- Arunajadai SG, Rauh VA. Handling covariates subject to limits of detection in regression. *Environmental and Ecological Statistics*. 2012; 157:369–391.
- Austin PC, Brunner LJ. Type I error inflation in the presence of a ceiling effect. *The American Statistician*. 2003; 57:97–104.
- Austin PC, Hoch JS. Estimating linear regression models in the presence of a censored independent variable. *Statistics in Medicine*. 2004; 23:411–429. [PubMed: 14748036]
- D'Angelo GD, Weissfeld L. An index approach for the cox model with left censored covariates. *Statistics in Medicine*. 2008; 27:4502–4514. [PubMed: 18407573]
- Doehler K, Davidian M. 'Smooth' inference for survival functions with arbitrarily censored data. *Statistics in Medicine*. 2008; 27:5421–5439. [PubMed: 18613273]
- Gallant AR, Nychka DW. Semi-nonparametric maximum likelihood estimation. *Econometrica*. 1987; 55:363–390.
- Giovanini, J. Ph.D. thesis. Oregon State University; 2008. Generalized linear mixed models with censored covariates.
- Helsel, DR. *Statistics for Censored Environmental Data Using Minitab and R*. Wiley. 2012.
- Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Applied Occupation and Environmental Hygiene*. 1990; 5:46–51.
- Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, Fine J, Krichevsky A, Delude R, Angus D. Understanding the inflammatory cytokine response in pneumonia and sepsis. *Archives of Internal Medicine*. 2007; 167:1655–1663. [PubMed: 17698689]
- Langohr K, Gomez G, Muga R. A parametric survival model with an interval-censored covariate. *Statistics in Medicine*. 2004; 23:309–319.
- Lee S, Park SH, Park J. The proportional hazards regression with a censored covariate. *Statistics and Probability Letters*. 2003; 61:309–319.
- Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1982; 44:226–233.
- Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environmental Health Perspectives*. 2004; 112:1691–1696. [PubMed: 15579415]

- Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine*. 2001; 20:33–45. [PubMed: 11135346]
- Martin AD, Quinn KM, Park JH. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*. 2011; 42:22.
- Rigobon R, Stoker TM. Estimation with censored regressors: basic issues. *International Economic Review*. 2007; 48:1441–1467.
- Rigobon R, Stoker TM. Bias from censored regressors. *Journal of Business and Economic Statistics*. 2009; 27:340–353.
- Robins JM, Wang N. Inference for imputation estimators. *Biometrika*. 2000; 87:113–124.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley; 1987.
- Sattar A, Sinha SK, Morris NJ. A parametric survival model when a covariate is subject to left-censoring. *Journal of Biometrics and Biostatistics*. 2012; S3:002. doi:10.4172/2155–6180.S3–002.
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
- Tang, L. Ph.D. thesis. North Carolina State University; 2008. ‘Smooth’ inference for clustered survival data.
- Therneau, T. A Package for Survival Analysis in S. R package version 2.36-12. 2012.
- Tsiatis, AA. *Semiparametric Theory and Missing Data*. Springer; 2006.
- Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. *Biometrika*. 1998; 84:935–948.
- Zhang D, Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*. 2001; 57:795–802. [PubMed: 11550930]
- Zhang M, Davidian M. ‘Smooth’ semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*. 2007; 64:567–576. [PubMed: 17970813]

Figure 1.

Normal Q-Q plots for $\log(\text{TNF})$, $\log(\text{IL-6})$, and $\log(\text{IL-10})$. The Y-axis corresponds to the sorted, non-censored, centered and scaled log-cytokine values. The X-axis corresponds to the quantiles of a $\mathcal{N}(0, 1)$ above the p_0 th quantile.

Table 1

Simulation results for an AFT regression model with Weibull errors, left, and a mixture of Weibull and lognormal errors, right, for 1000 datasets.

Method		Weibull Errors				Mixture Errors			
		0	1	2	3	0	1	2	3
Weibull CC	Bias	-0.001	-0.002	-0.009	0.000	2.050	-0.018	-0.007	-0.002
	SD	0.847	0.140	0.110	0.009	2.315	0.383	0.303	0.026
	SE	0.850	0.144	0.114	0.009	1.400	0.250	0.214	0.017
Parametric CC	Bias	-0.038	-0.002	-0.009	0.001	-0.008	0.001	-0.002	-0.002
	SD	0.865	0.142	0.112	0.009	1.400	0.231	0.180	0.015
	SE	0.855	0.139	0.110	0.009	1.423	0.262	0.201	0.010
SNP CC	Bias	-0.082	-0.002	-0.008	0.001	-0.055	-0.000	-0.000	0.001
	SD	0.939	0.144	0.113	0.010	1.372	0.163	0.129	0.012
	SE	0.931	0.141	0.115	0.011	1.340	0.161	0.125	0.014
Weibull SUB	Bias	5.572	-0.285	-1.551	0.038	15.110	-2.460	-1.382	0.028
	SD	1.043	0.236	0.178	0.013	2.064	0.300	0.252	0.028
	SE	0.984	0.154	0.104	0.012	1.550	0.242	0.207	0.020
Parametric SUB	Bias	4.384	-0.314	-1.332	0.021	8.595	-1.340	-0.712	0.004
	SD	0.930	0.190	0.257	0.011	1.375	0.223	0.170	0.016
	SE	0.931	0.151	0.131	0.011	1.292	0.219	0.175	0.016
SNP SUB	Bias	5.888	-0.088	-0.210	0.001	12.778	-0.451	0.030	0.001
	SD	0.753	0.084	0.179	0.007	1.214	0.178	0.137	0.012
	SE	0.654	0.089	0.079	0.007	1.050	0.168	0.132	0.012
Weibull MI	Bias	-0.050	-0.000	0.003	0.001	1.191	-0.036	0.043	-0.001
	SD	0.430	0.066	0.059	0.006	1.180	0.229	0.177	0.016
	SE	0.438	0.065	0.059	0.006	0.892	0.152	0.137	0.012
Parametric MI	Bias	-0.077	-0.000	0.003	0.001	-0.001	-0.001	0.007	-0.002
	SD	0.461	0.067	0.060	0.006	0.710	0.112	0.077	0.009
	SE	0.440	0.066	0.059	0.005	0.714	0.116	0.087	0.010
SNP MI	Bias	-0.035	-0.000	0.004	0.001	-0.054	-0.014	0.009	-0.001
	SD	0.470	0.068	0.062	0.006	0.544	0.094	0.068	0.007
	SE	0.469	0.066	0.061	0.006	0.522	0.092	0.070	0.007
Weibull IMI	Bias	0.002	0.000	0.002	0.000	4.307	-0.520	0.050	0.000
	SD	0.417	0.065	0.052	0.006	1.239	0.225	0.170	0.017
	SE	0.431	0.063	0.050	0.006	1.009	0.196	0.176	0.014
Parametric IMI	Bias	-0.022	0.000	0.002	0.000	2.440	-0.008	-0.010	-0.001
	SD	0.440	0.065	0.052	0.006	0.719	0.112	0.076	0.009

		Weibull Errors				Mixture Errors			
		0	1	2	3	0	1	2	3
SNP IMI	SE	0.432	0.063	0.050	0.006	0.708	0.113	0.081	0.010
	Bias	−0.031	0.001	0.003	0.000	0.019	−0.010	0.003	0.000
	SD	0.430	0.065	0.052	0.006	0.504	0.087	0.056	0.006
	SE	0.430	0.064	0.050	0.006	0.490	0.085	0.056	0.006

SD: Monte Carlo standard deviation; SE: average estimated standard error; CC: complete case; SUB: substitution by DL/2; MI: multiple imputation; IMI: iterated multiple imputation

Table 2

Parameter estimates, standard error estimates, and p-values for coefficients of log-cytokine and demographic covariates in the AFT model for survival time, based on the GenIMS dataset.

	CC			MI			IMI		
	Par.	SE	p-value	Par.	SE	p-value	Par.	SE	p-value
Int.	13.491	1.804	<0.001	15.351	0.995	<0.001	14.891	0.921	<0.001
TNF	-0.460	0.237	0.052	-0.202	0.138	0.143	-0.189	0.134	0.159
IL-6	-0.138	0.099	0.164	-0.122	0.071	0.084	-0.120	0.070	0.091
IL-10	0.005	0.168	0.977	-0.125	0.097	0.196	-0.144	0.094	0.097
Sex	-0.488	0.357	0.172	-0.540	0.249	0.030	-0.529	0.248	0.028
Race	-0.017	0.647	0.980	-0.254	0.409	0.504	-0.252	0.407	0.536
Age	-0.084	0.016	<0.001	-0.074	0.010	<0.001	-0.074	0.010	<0.001

Par.: parameter estimate; SE: standard error estimate; CC: complete case; MI: multiple imputation; IMI: iterated multiple imputation