Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Sattar et al., J Biomet Biostat 2012, S3
DOI: 10.4172/2155-6180.S3-002

Research Article Open Access

A Parametric Survival Model When a Covariate is Subject to Left-Censoring

Abdus Sattar[1]*, Sanjoy K. Sinha[2] and Nathan J. Morris[1]

[1]*Department of Epidemiology & Biostatistics, Case Western Reserve University, Cleveland, OH, USA*
[2]*School of Mathematics and Statistics, Carleton University, Ottawa, Ontario K1S 5B6, Canada*

Abstract

Problem statement: Modeling survival data with a set of covariates usually assumes that the values of the covariates are fully observed. However, in a variety of applications, some values of a covariate may be left-censored due to inadequate instrument sensitivity to quantify the biospecimen. When data are left-censored, the true values are missing but are known to be smaller than the detection limit. The most commonly used ad-hoc method to deal with nondetect values is to substitute the nondetect values by the detection limit. Such ad-hoc analysis of survival data with an explanatory variable subject to left-censoring may provide biased and inefficient estimators of hazard ratios and survivor functions.

Method: We consider a parametric proportional hazards model to analyze time-to-event data. We propose a likelihood method for the estimation and inference of model parameters. In this likelihood approach, instead of replacing the nondetect values by the detection limit, we adopt a numerical integration technique to evaluate the observed data likelihood in the presence of a left-censored covariate. Monte Carlo simulations were used to demonstrate various properties of the proposed regression estimators including the consistency and efficiency.

Results: The simulation study shows that the proposed likelihood approach provides approximately unbiased estimators of the model parameters. The proposed method also provides estimators that are more efficient than those obtained under the ad-hoc method. Also, unlike the ad-hoc estimators, the coverage probabilities of the proposed estimators are at their nominal level. Analysis of a large cohort study, genetic and inflammatory marker of sepsis study, shows discernibly different results based on the proposed method.

Conclusion: Naive use of detection limit in a parametric survival model may provide biased and inefficient estimators of hazard ratios and survivor functions. The proposed likelihood approach provides approximately unbiased and efficient estimators of hazard ratios and survivor functions.

Keywords: Left-censored covariate; Maximum likelihood method; Numerical integration; Survival model

Introduction

Survival models are commonly used to assess the relationship between a covariate of interest and time-to-event data. In these models it is typically assumed that the covariate is fully observed, but there are many situations when the underlying covariate is not fully observed. Incomplete measurements of a variable can occur in environmental, epidemiological, biological and biomedical studies [1-3]. For example, when conducting a bioassay to quantify the biomarker some measurements are not fully observed because of inadequate instrument sensitivity. Similar incomplete measurements can also occur when measuring air quality, water quality, soils, contaminants in biota, etc. The measurement above the detection limit (LOD) is reported, and in those that are undetectable, LOD is reported. Several authors [2] reported that the use of the LOD or LOD/2 provide biased regression parameter estimate. When studying an association between a biomarker subject to LOD and time-to-event, it is necessary to adjust the impact of LOD in survival analysis. In this article we intend to study the association between a right censored survival outcome and a left-censored covariate based on the direct maximization of a likelihood function where the impact of left-censoring in the covariate of interest will be integrate out by a numerical integration method.

As a running example, we use the Genetic and Inflammatory Marker of Sepsis (GenIMS) study. This was a large cohort study of patients with community-acquired pneumonia and sepsis [4]. The goal of this study was to understand the role of inflammatory cytokine response in a hospitalized cohort of patients. After enrollment in the study, blood was drawn for a cytokine assay immediately following the enrollment, daily for the first week and weekly thereafter while subjects remained in the hospital. There are several cytokine measured in this study and one of them is Interleukin 10 (IL10). About one-third of the IL10 measurements fall below the detection limit and LOD is reported. In this case IL10 is a risk factor or a covariate of interest which has left-censoring. Our goal is to find the association between 90 day mortality and IL10 given that a large percentage of IL10 measurements are left-censored. More details about the GenIMS study can be found in the result section.

During the past several years new methods have been developed for improved statistical inference when there is a censored covariate in the regression model, and these methods have been compared with naive methods. Naive methods include removing observations falling below the detection limit. Removing observations may provide unbiased regression parameter estimates but results in reduced sample size

***Corresponding author:** Abdus Sattar, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, 10900 Euclid Avenue, BRB, G-19, Cleveland, OH 44106-4945, USA, Tel: 1.216.368.1501; Fax: 1.216.368.1969; E-mail: sattar@case.edu

Received July 05, 2012; **Accepted** August 20, 2012; **Published** August 25, 2012

Citation: Sattar A, Sinha SK, Morris NJ (2012) A Parametric Survival Model When a Covariate is Subject to Left-Censoring. J Biomet Biostat S3:002. doi:10.4172/2155-6180.S3-002

and hence decreasing efficiency of the parameter estimates. Another commonly practiced approach is the ad-hoc substitution method. In this approach observations that fall below the detection limit are recognized by LOD, LOD/2, LOD/ $\sqrt{2}$, or zero. Helsel [5] and Sattar et al. [6] studied these ad-hoc methods and showed that these ad-hoc methods provide biased estimates and the degree of bias increases with the increase in percent of LOD observations in the covariate. Helsel argued that there is no theoretical basis for the use of these substitution methods. Two articles on censored covariate in the generalized linear model appeared almost at the same time in the literature, one used a maximum likelihood method with the Monte Carlo EM algorithm (May et al. [7]), and the other used an optimal estimating equations approach (Tsimikas et al. [8]). Nie et al. [9] studied left-censoring of an explanatory variable in the linear regression model set-up. These authors demonstrated that the commonly used substitution methods of replacing left-censored values with LOD, LOD/$\sqrt{2}$, LOD/2 provide biased parameter estimates with low Coverage Probabilities (CP). They proposed parameter estimation by the maximum likelihood method based on parametric distributional assumptions. The proposed method has been compared with a method of replacing LOD by E(X|X<LOD). The authors concluded that these two methods are competitive and are promising alternatives to the multiple imputation method [10].

There are several approaches to model the hazard of an event. A common approach is the parametric survival model. In this type of modeling, a probability distribution is assumed for the underlying survival time. If the distributional assumption is satisfied then this modeling approach is more efficient than its counterpart nonparametric and semi-parametric hazard models. Langor et al. [11] studied doubly censored survival data with an interval-censored covariate in a parametric survival model framework. They have considered a censored discrete covariate. In their estimation approach, the likelihood function is maximized as a non-linear constant maximization problem, and they used a sequential quadratic programming algorithm. This approach guarantees a local maximum likelihood estimate. Cox regression models with covariate subject to detection limit has also been studied. Lee et al. [12] propose to estimate the relative risk function based on the uncensored covariate data and used this risk function to derive a partial likelihood function. D'Angelo et al. [13] analyzed survival data in a Cox model framework with a covariate subject to left-censoring. These authors have used an index approach which is conceptually similar to the EM algorithm. In this approach the censored value is expressed as a function of all of the observed values of the covariate.

In this article, we propose a method for estimating survival regression parameter associated with a continuous covariate of interest which is subject to limit of detection. The covariate of interest is left-censored because of the limit of detection in the bioassay. We maximized the likelihood function and integrate out the left-censoring via Simpson's numerical integration method. Monte Carlo simulations study show that the proposed method provides approximately unbiased estimates of the model parameters. The parameter estimates are also efficient and its Coverage Probabilities (CP) is at the nominal level. The method has been implemented in standard statistical software. To our knowledge, no one has addressed the detection limit problem in a parametric survival model using a numerical integration method.

The article is organized as follows, in section "Materials and Methods", we have developed the general framework of our proposed method. In sections "Simulation Study" and "Illustrative Example", we have presented the simulation and GenIMS study results, respectively. We have offered a discussion and conclusion in the final section.

Materials and Methods

Suppose in an experiment with n subjects, T_i denotes the survival time of subject i, i=1,…,n. Assume that some of the "true" values $t_1,t_2,\ldots,t_n$ of the random variables $T_1,\ldots,T_n$ are right-censored. We further assume that the censoring is non-informative. The right-censored observed survival data can be written as pairs (y_i,δ_i), where δ_i is the event indicator: δ_i=1 if y_i is a true event time, that is, if t_i= y_i, and δ_i=0 if t_i is right-censored, that is, if t_i> y_i. Let X_i denote a p×1 vector of covariates associated with the i^{th} subject. Suppose the hazard rate $h_i(t)$ for subject i at time t is related to the values x_i of the covariates by the proportional hazards model

$$h_i(t) = exp(x_i'\beta_1)h_0(t)$$

where $h_0(t)$ is a baseline hazard function depending on unknown parameters β_0 and β_1 is a p×1 vector of unknown regression coefficients. Assuming that the survival times are independent, the likelihood function of $\beta = (\beta_0, \beta_1)$ for given data (y_i,δ_i,x_i) can be defined as

$$L_0(\beta) = \prod_{i=1}^{n} \{h_i(y_i)\}^{\delta_i} S_i(y_i),$$

where $S_i(t) = P(T_i > t \mid x_i, \beta)$ is the survivor function for subject i at time t. Let $f_{T_i}(t \mid x_i, \beta)$ denote the density function for the survival time T_i at time t. Then the above likelihood function can be expressed as

$$L_0(\beta) = \prod_{i=1}^{n} \{f_{T_i}(y_i \mid x_i, \beta)\}^{\delta_i} \{P(T_i > y_i \mid x_i, \beta)\}^{1-\delta_i}. \quad (1)$$

When the values of a covariate are censored due to the limit of detection, and the censored values are replaced by the LOD, then likelihood function (1) provides biased and inefficient regression parameter estimates [13,14]. To obtain consistent and efficient regression parameter estimates from a survival regression model with a covariate subject to left-censoring we are proposing the following method. This method is based on Simpson's numerical integration technique and easy to implement in standard statistical software. The likelihood function can be constructed for the censored and observed values with a fair amount of effort. For now we consider that X_i has only one continuous covariate and its value x_i is left-censored if x_i < C for a given value of c. Let $f_{X_i}(x)$ denote the density of the random variable X_i, which is assumed to be known. Define a binary random variable R_i which is 1 if X_i is observed and 0 if X_i is not detected, that is,

$$R_i = \begin{cases} 1 & for \quad X_i \geq c \\ 0 & for \quad X_i < c \end{cases}$$

We assume that the binary random variable R_i follows the Bernoulli distribution

$$f_{R_i}(r) = \pi_i^r (1-\pi_i)^{1-r}$$

for r = 0,1, where $\pi_i = P(X_i \geq c)$ is the probability that the value of X_i is observed. To estimate the model parameters β, we propose to maximize the observed data likelihood function

$$L(\beta) = \prod_{i=1}^{n} \left\{ \{f_{T_i}(y_i \mid x_i, \beta)\}^{\delta_i} \{P(T_i > y_i \mid x_i, \beta)\}^{1-\delta_i} f_{R_i}(r_i) f_{X_i}(x_i) \right\}^{r_i}$$
$$\times \left\{ \int_{-\infty}^{c} \{f_{T_i}(y_i \mid x_i, \beta)\}^{\delta_i} \{P(T_i > y_i \mid x_i, \beta)^{1-\delta_i} f_{R_i}(r_i) f_{X_i}(x_i) dx_i \right\}^{1-r_i}. \quad (2)$$

In the absence of left-censored covariates, the above likelihood function $L(\beta)$ becomes the ordinary likelihood $L_0(\beta)$, as defined in (1). From (2), the log-likelihood function is obtained as

$$l(\beta) = \sum_{i=1}^{n} r_i \log \left\{ \{f_{T_i}(y_i \mid x_i, \beta)\}^{\delta_i} \{P(T_i > y_i \mid x_i, \beta)\}^{1-\delta_i} f_{R_i}(r_i) f_{X_i}(x_i) \right\}$$
$$+ \sum_{i=1}^{n} (1-r_i) \log \left\{ \int_{-\infty}^{c} \{f_{T_i}(y_i \mid x_i, \beta)\}^{\delta_i} \{P(T_i > y_i \mid x_i, \beta)^{1-\delta_i} f_{R_i}(r_i) f_{X_i}(x_i) dx_i \right\}. \quad (3)$$

Note that the above log-likelihood function (3) cannot be written in a closed form, and numerical methods may be used to evaluate the integral with respect to the covariate x_i. Here we consider evaluating this integral using Simpson's 1/3 rule of numerical integration. The Simpson's method produces a numerical value for the integration of a function over a set. Suppose that an interval [a,b] is divided into k subintervals, with k an even number. Then the composite Simpson's rule is defined by [15]

$$\int_a^b f(z)dz \approx \frac{h}{3}\left[f(z_0) + 2\sum_{j=1}^{n/2-1} f(z_{2j}) + 4\sum_{j=1}^{n/2} f(z_{2j-1}) + f(z_n) \right],$$

where $z_j = a + jh$ for $j = 0,1,\ldots,n$, with $h = (b\text{-}a)/n$. The error term associated with the composite Simpson's rule is bounded (in absolute value) by $(h^4/180)(b-a)\max_{\xi\in[a,b]}\left|f^4(\xi)\right|$. Differentiating $l(\beta)$ with respect to β, gives the score equations $U(\beta) = (\partial/\partial\beta)l(\beta) = 0$. The maximum likelihood estimators of the model parameters β can be obtained by solving these score equations numerically using an iterative method or by directly maximizing the log-likelihood function (3) using some numerical optimization technique, which is discussed further in the next section.

Standard maximum likelihood theory suggests that $E\{U(\beta)\} = 0$. The observed Fisher information $I(\beta)$ is the negative of the $p\times p$ Hessian matrix of the log-likelihood, so that $I(\beta) = -(\partial^2/\partial\beta\partial\beta')l(\beta) = -(\partial/\partial\beta)U(\beta)$. For the exponential family, the expected Fisher information matrix, $J(\beta) = E\{U(\beta)U'(\beta)\} = -E\{(\partial/\partial\beta)\ U(\beta)\}$. Under appropriate regularity conditions, the maximum likelihood estimators follow an approximate normal distribution for a large sample size n:

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, J^{-1}(\beta)).$$

Simulation study

To examine the performance of the proposed method, we conducted a simulation study. In this study, we compared our proposed method based on the log-likelihood function (3) with the naïve method, which estimates the model parameters by replacing the left-censored covariates with the LOD under a number of different scenarios. We refer to these two methods of analysis as the "corrected"

Scenario			Intercept(β_0)				Slope(β_1)		
N	SD of X	LOD	Analysis	Bias	MSE	CP	Bias	MSE	CP
500	0.5	0%	Naïve	0.004	0.300	94.5%	-0.00014	0.0118	95.1%
			Corrected	0.004	0.300	94.5%	-0.00014	0.0118	95.1%
		20%	Naïve	0.139	0.451	94.6%	-0.02549	0.0174	94.8%
			Corrected	-0.021	0.312	94.9%	0.004033	0.0124	94.7%
		50%	Naïve	0.574	1.309	90.5%	-0.10275	0.0466	91.2%
			Corrected	0.027	0.376	94.8%	-0.00509	0.0148	94.8%
	1	0%	Naïve	0.010	0.069	95.4%	-0.00148	0.0027	95.9%
			Corrected	0.010	0.069	95.4%	-0.00148	0.0027	95.9%
		20%	Naïve	0.190	0.150	91.6%	-0.03297	0.0055	92.0%
			Corrected	0.010	0.078	94.6%	-0.00185	0.0031	95.1%
		50%	Naïve	0.547	0.571	81.4%	-0.08614	0.0167	85.7%
			Corrected	-0.019	0.098	95.0%	0.004916	0.0037	95.5%
	2	0%	Naïve	0.002	0.019	95.9%	-0.00017	0.0007	95.5%
			Corrected	0.002	0.019	95.9%	-0.00017	0.0007	95.5%
		20%	Naïve	0.212	0.080	78.6%	-0.03273	0.0023	83.5%
			Corrected	0.007	0.023	93.5%	-0.00102	0.0009	94.5%
		50%	Naïve	0.724	0.612	33.7%	-0.09892	0.0125	52.4%
			Corrected	0.012	0.029	95.4%	-0.00127	0.0011	94.9%
1000	0.5	0%	Naïve	0.005	0.148	94.6%	-0.00101	0.0058	95.2%
			Corrected	0.005	0.148	94.6%	-0.00101	0.0058	95.2%
		20%	Naïve	0.168	0.247	93.3%	-0.03074	0.0095	93.5%
			Corrected	0.009	0.158	94.4%	-0.00151	0.0063	94.0%
		50%	Naïve	0.530	0.781	87.1%	-0.09401	0.0274	87.9%
			Corrected	0.008	0.189	94.6%	-0.00100	0.0075	94.4%
	1	0%	Naïve	-0.002	0.038	94.5%	0.000437	0.0015	94.5%
			Corrected	-0.002	0.038	94.5%	0.000436	0.0015	94.5%
		20%	Naïve	0.188	0.090	88.0%	-0.03253	0.0032	89.9%
			Corrected	0.008	0.038	95.3%	-0.00152	0.0015	95.5%
		50%	Naïve	0.588	0.491	65.3%	-0.09484	0.0140	70.1%
			Corrected	0.003	0.052	94.2%	-0.00059	0.0020	94.3%
	2	0%	Naïve	-0.001	0.010	95.8%	-0.00015	0.0004	95.0%
			Corrected	-0.001	0.010	95.8%	-0.00016	0.0004	95.0%
		20%	Naïve	0.210	0.059	62.9%	-0.03248	0.0016	72.9%
			Corrected	0.003	0.010	95.5%	-0.00055	0.0004	96.1%
		50%	Naïve	0.724	0.569	6.7%	-0.09928	0.0112	21.9%
			Corrected	0.008	0.015	95.1%	-0.00105	0.0005	95.8%

Table 1: Simulation results of a parametric survival model when a covariate is subject to limit of detection. The true values of the regression parameters set to $\beta_0 = -2.0$, and $\beta_1 = -0.2$.

and the "naïve" approach, respectively. In each scenario, we consider a Weibull proportional hazard model. Under this proportional hazards model, the hazard of death at time t for the i^{th} subject is [16]

$$h_i(t) = exp(x_i'\beta_1)\lambda\gamma t^{\gamma-1} \quad (4)$$

where λ and γ are the scale and shape parameters of the Weibull distribution, respectively. The survivor function corresponding to the hazard function (4) is $S_i(t) = exp\left\{-exp\left(x_i'\beta\right)\lambda t^{\gamma}\right\}$. For simplicity, we set the shape parameter γ =1. In this setting, the hazard function (4) can be written in the form $h_i(t) = exp(x_i^{*\prime}\beta)$, where $x_i^{*} = (1, x_i)'$ and $\beta = [\beta_0, \beta_1]'$ with $\beta_0 = \log(\lambda)$. The values of the covariate X were generated from the normal distribution with mean 5.0 and standard deviation which differed for some of the scenarios. True values of the regression parameters intercept (β_0) and slope (β_1) were set to -2.0 and -0.2, respectively. The right-censored survival times were generated from the Weibull distribution by setting λ = exp(intercept + 50) If the observed time is less than the right-censoring time, then the event is observed. Otherwise, the survival time is right-censored. The values that differed for each scenario were the sample size ($N \in \{500, 1000\}$), the standard deviation of the covariate ($SD(X_1) \in \{0.5, 1.0, 2.0\}$) and the percentages of covariates which were censored $(1-\pi \in \{0\%, 20\%, 50\%\})$.To generate various percentages of left-censored covariate values, we set LOD = 5+SD(X_1)Φ^{-1}(1-π), where Φ is the normal cumulative density function. If the generated values of the covariate X_1 are less than the LOD, then LOD is recorded. The statistical software R [13] was used for the computation. In particular, to maximize the likelihood function derived under the above Weibull proportional hazard model, we used the method L-BFGS-B [14] available through the R function optim. This method uses function values and gradients to build up a picture of the surface to be optimized. For the naive approach we used the survival package in R. The simulation results are presented in Table 1. As expected with no LOD (i.e. 1-π = 0%), the naïve approach and corrected approach are identical. As the proportion of censored values increased, the bias in the estimates from the naïve approach also increased. Also, the bias in the estimates from the naïve approach was significantly higher when the standard deviation of the covariate was higher. When the standard deviation of the covariate was 2.0 with a sample size 1000 and 50% observations were left-censored, the estimated 95% coverage rate for both the slope and intercept was less than 22% for the naïve approach. In contrast, the corrected approach produced results with very small bias, smaller mean square error, and approximately correct coverage for most scenarios. When the variance of the covariate was 2.0, the corrected approach had a slightly low coverage rate for 500 sample size, but significantly improved coverage compared to the naïve approach. Thus the proposed approach is approximately unbiased and achieves good coverage rates in most of the scenarios.

Illustrative example

Severe sepsis is the systemic inflammatory response to infection with complication of organ dysfunction. Community-Acquired Pneumonia (CAP) is the leading cause of severe sepsis. The Genetic and Inflammatory Markers of Sepsis (GenIMS) study - a large, multicenter, cohort study of patients with CAP was conducted to understand the pattern of systemic cytokine response to infection and to determine if there were specific patterns associated with severe sepsis and death [17]. A total of 2320 patients with CAP presenting to the emergency departments of 28 hospitals in Pennsylvania, Connecticut, Michigan, and Tennessee enrolled in the study during December 2001 and November 2003. GenIMS included patients with age ≥ 18 years old with a clinical and radiologic diagnosis of pneumonia. After enrollment detail baseline and clinical information were gathered, and blood was drawn for cytokine assays immediately following enrollment and daily throughout the first seven days of hospitalization. The primary outcome variable in the GenIMS study was severe sepsis and 90-day mortality. The markers of greatest interest in the GenIMS study were the pro-inflammatory marker Interleukin-6 (IL6) and anti-inflammatory marker Interleukin-10 (IL10). More information regarding the study population, outcomes, treatment, and covariates can be found in the Kellum et al. [17].

In this illustration, we consider the association between 90-day mortality and the IL10 biomarker baseline (Day 1) data. Blood was drawn for a cytokine assay from 1429 subjects. If the patients presented to the emergency department after 11 pm or on the weekends or holidays, then the blood was not drawn for logistic reasons. Note that there are some intermittent missing biomarker data due to administrative reasons and we are assuming that this intermittently missing data are missing completely at random. A detail decomposition of the study subjects can be found in the above mentioned reference. In this analysis, we have a total 867 subjects with IL10 measurements at baseline. However, the measurements of IL10 were left-censored (47.87 percent) because of the inadequate sensitivity of the cytokine assay resulting in a left-censoring of the measure at the lower limit of detection.

Table 2 reports the descriptive statistics of the covariates that we consider in this analysis. The presented result is based on the baseline (Day 1) characteristics of demographic and clinical variables. From this table we can say that the patients who have died during the first 90 days after the hospitalization for CAP were mostly male and older patients. Higher proportions of these patients had been treated with steroids, and their D-dimer and IL10 levels were higher.

Table 3 summarizes the results from the GenIMS data analyses. To examine the impact of left-censoring in a real study, we have fit the corrected and naïve models described in the simulation section. The naïve survival model is a parametric Weibull survival model where nondetect values are replaced by the LOD. The corrected survival model is our proposed model where we have fitted the survival model with an implementation of the Simpson's numerical integration technique for the left-censoring for IL10. The model considered includes the anti-inflammatory biomarker IL10, age, gender, steroid use, and coagulation marker D-dimer. We have performed the logarithmic transformation on the continuous skewed data (IL10 and D-dimer), and rescale the age variable (age ÷ 10) so that the estimates become stable and have improved the interpretation. The estimates from the two models are different. The corrected model Hazard Ratio (HR) estimate for the covariate IL10 is smaller than its counterpart naive model HR estimate. The proposed model HR estimate for IL10 is also more efficient than the other model. The 95% CI of the HR estimate for IL10 obtained from

Variable	Sample size (survivors)	Survivors (Mean ± SD)	Sample size (nonsurvivors)	Nonsurvivors (Mean ± SD)
Age	789	64.64 ± 17.86	86	78.53 ± 11.46
Gender(Female)	789	0.49 ± 0.02	86	0.41 ± 0.05
Steroid[1]	789	0.35 ± 0.02	86	0.43 ± 0.05
D-dimer[1]	787	6.02 ± 1.59	86	6.83 ± 1.52
IL10[2]	782	2.28 ± 1.0	85	2.69 ± 1.21

[1]Log transformation is applied to D-dimer and IL10; [2]Censored values are replaced by the limit of detection (LOD) 5.

Table 2: Demographic and clinical characteristics of survivors and nonsurvivors at baseline.

Model	Naive Survival Model				Corrected Survival Model					
Covariate	Coefficient	SE	P-val	95% CI	Coefficient	SE	P-val	95% CI		
				LL[4]	UL[5]				LL	UL
IL10[1]	0.267	0.084	0.001	0.103	0.431	0.232	0.065	0.000	0.105	0.359
Age[2]	0.593	0.098	0.000	0.401	0.786	0.602	0.099	0.000	0.409	0.796
Gender	-0.431	0.224	0.054	-0.870	0.008	-0.436	0.224	0.051	-0.874	0.003
Steroid	0.393	0.222	0.076	-0.042	0.828	0.351	0.223	0.115	-0.086	0.788
D-dimer[3]	0.281	0.093	0.003	0.099	0.463	0.276	0.093	0.003	0.094	0.458

[1]Log of IL10 where censored values are replaced by the LOD=5; [2]Age is divided by 10 for better interpretation; [3]Log of D-dimer; [4]LL denotes lower limit of the 95% confidence interval; [5]UL denotes the upper limit of the 95% confidence interval.

Table 3: GenIMS study results of a parametric survival model when IL10 is subject to limit of detection.

the naive and corrected models are [1.108, 1.539] and [1.111, 1.432] respectively. These results suggest that the naive use of the detection limit as a substitution for an undetected value can lead to a different estimate and interpretation of the risk factors. Our simulation results have shown that there are situation where the difference between the two approaches is even larger than in our real data example.

Discussion

A censored covariate is a challenge for statistical analysis. We consider left-censoring of a covariate and examined the impact of left-censoring in a parametric survival model. There are several ad-hoc methods to deal with the limit of detection problem of a covariate in a regression model framework. These methods provide biased and inefficient parameter estimates. In this paper we proposed a method for correcting bias and making an efficient parametric survival inference when there is a left-censored covariate. Our propose likelihood method is based on Simpson's numerical integration technique. Because the data involves both a right-censored time-to-event outcome and a left-censored covariate, the likelihood function becomes a complicated one. From this complicated likelihood function, we have integrated out the impact of left-censoring. The Monte Carlo simulation study shows that the proposed model's performance is comparable to the standard survival model's performance where there is no left-censoring. We have also applied the proposed method to a large cohort data set. From this exercise we have found that the proposed method results are different from the ad-hoc method results.

In the situation when a covariate is subject to left-censoring, this paper compares a new method for analyzing survival data to a commonly used naive method that replace the censored values by the limit of detection. We have demonstrated that the naive method provide biased, efficient regression parameter estimates with low coverage probabilities. On the other hand our proposed likelihood method based on a numerical integration technique provides approximately unbiased and efficient parameter estimates, and achieves good coverage probabilities in most of the scenarios. The proposed method is relatively simple to understand and easy to implement in a standard statistical software.

We have implemented our proposed method by considering only one covariate with limit of detection. We expect that this method can be extended with some computational burden for more than one covariate with the limit of detection. A limitation of this study is that we assumed a normal distribution for the censored covariate and derive the likelihood function accordingly, and we did not investigate the robustness to the misspecification of the normality assumption in the simulation. We also did not examine the impact of changing the shape parameter value for the Weibull distribution in our simulation. We are working on another manuscript where we are intending to relax the assumption of normality, and perform sensitivity analysis. In summary, in the presence of limit of detection in a covariate of a parametric survival model, the estimates are biased and inefficient. Our proposed likelihood-based method using a numerical integration provides unbiased and efficient parameter estimates. Therefore, the proposed method is an encouraging one to use when a covariate is subject to a limit of detection. The statistical analysis was performed using R software version 2.15.0. The R script can be obtained upon request to the corresponding author.

Acknowledgements

We thank Dr. Derek Angus and the CRISMA laboratory for access to the GenIMS data. We are indebted to the nurses, respiratory therapists, phlebotomists, physicians, and other health-care professionals, as well as the patients and their families who supported this trial. A complete list of GenIMS investigators is available at www.ccm.upmc.edu/genims investigators. The GenIMS study was funded via grant R01 GM61992 by the National Institute of General Medical Sciences. Individuals and institutions who participated in GeniMS study can be found in the Appendix I.

References

1. Helsel DR (2005) Nondetects and data analysis: statistics for censored environmental data. Wiley-Interscience, USA.
2. Schisterman EF, Little RJ (2010) Opening the black box of biomarker measurement error. Epidemiology 21: S1-S3.
3. Cho JC, Tiedje JM (2002) Quantitative detection of microbial genes by using DNA microarrays. Appl Environ Microbiol 68: 1425-1430.
4. Yende S, D'Angelo G, Kellum JA, Weissfeld L, Fine J, et al. (2008) Inflammatory markers at hospital discharge predict subsequent mortality after pneumonia and sepsis. Am J Respir Crit Care Med 177: 1242-1247.
5. Helsel DR (2006) Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65: 2434-2439.
6. Sattar A, Weissfeld LA, Molenberghs G (2011) Analysis of non-ignorable missing and left-censored longitudinal data using a weighted random effects tobit model. Stat Med 30: 3167-3180.
7. May RC, Ibrahim JG, Chu H (2011) Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. Stat Med 30: 2551-2561.
8. Tsimikas JV, Bantis LE, Georgiou SD (2012) Inference in generalized linear regression models with a censored covariate. Comput Stat Data Anal 56: 1854-1868.
9. Nie L, Chu H, Liu C, Cole SR, Vexler A, et al. (2010) Linear regression with an independent variable subject to a detection limit. Epidemiology 21: S17-S24.
10. Little RJA, Rubin DB (2002) Statistical analysis with missing data. Wiley: New York.
11. Langohr K, Gómez G, Muga R (2004) A parametric survival model with an interval-censored covariate. Stat Med 23: 3159-3175.
12. Lee S, Park SH, Park J (2003) The proportional hazards regression with a censored covariate. Stat Probab Lett 61: 309-319.

13. D'Angelo G, Weissfeld L (2008) An index approach for the Cox model with left censored covariates. Stat Med 27: 4502-4514.

14. Wu L (2010) Mixed effects models for complex data. CRC Press: New York.

15. Cheney W, Kincaid D (2011) Numerical Analysis: Mathematics of Scientific Computing (3rd edition edn). Brooks/Cole, Thomson Learning: Pacific Grove.

16. Collett D (2003) Modelling Survival Data in Medical Research (2nd edn) Chapman & Hall/CRC: New York.

17. Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, et al. (2007) Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. Arch Intern Med 167: 1655-1663.

This article was originally published in a special issue, **Data analysis: Missing and multiway** handled by Editor(s). Dr. Keumhee Chough Carriere, University of Alberta, Canada.